Lecture 5 : Sparse Models

• Homework 3 discussion (Nima)

• Sparse Models Lecture
  – Reading : Murphy, Chapter 13.1, 13.3, 13.6.1
  – Reading : Peter Knee, Chapter 2

• Paolo Gabriel (TA) : Neural Brain Control

• After class
  – Project groups (Nima)
  – Installation Tensorflow, Python, Jupyter (TAs)
Homework 3: Fisher Discriminant

\[
P(c, 1x) = \frac{P(x | c_1) P(c_1)}{P(x)}
\]

\[
P(c_1 | x) = P(c_2 | x)
\]

\[
P(x | c_1) P(c_1) = P(x | c_2) P(c_2)
\]

\[
\Sigma = \Sigma_{c_1} = \Sigma_{c_2} \quad \hat{\mu}_1, \hat{\mu}_2
\]

\[
(x - \hat{\mu}_1)^T \Sigma_{c_1}^{-1} (x - \hat{\mu}_1) = (x - \hat{\mu}_2)^T \Sigma_{c_2}^{-1} (x - \hat{\mu}_2)
\]

\[
\text{trace} \left( \Sigma_{c_1}^{-1} \left[ (x - \hat{\mu}_1)^T (x - \hat{\mu}_1) - (x - \hat{\mu}_2)^T (x - \hat{\mu}_2) \right] \right) = 0
\]
\[ \begin{align*}
&+ \left( \mathbf{X}^{-1} \begin{bmatrix} x^T x - x^T \mu_1 - \mu_1^T x + \mu_1^T \mu_1 & x^T \mu_2 + \mu_2^T x - \mu_2^T \mu_2 \end{bmatrix} \right) = 0 \\
&+ \left( \mathbf{X}^{-1} \begin{bmatrix} 2 x^T (\mu_2 - \mu_1) + (\mu_2 - \mu_1)^T (\mu_2 + \mu_1) \end{bmatrix} \right) = 0 \\
&+ \left( \mathbf{X}^{-1} (x - \mu_2)^T (\mu_2 - \mu_1) \right) = 0 \\
&+ \left( \mathbf{X}^{-1} (2 x - (\mu_2 + \mu_1)) \right) = 0 \\
&+ \left( \mathbf{X}^{-1} (x - \mu_2) \right) = 0 \\
&+ \left( \mathbf{X}^{-1} (x - x_0) \right) = 0
\end{align*} \]
Sparse model

- Linear regression (with sparsity constraints)
- Slide 4 from Lecture 4

**Linear regression:** Linear Basis Function Models (1)

Generally

\[ y(x, w) = \sum_{j=0}^{M-1} w_j \phi_j(x) = w^T \phi(x) \]

- where \( \phi_j(x) \) are known as *basis functions*.
- Typically, \( \phi_0(x) = 1 \), so that \( w_0 \) acts as a bias.
- Simplest case is linear basis functions: \( \phi_d(x) = x_d \).

\[ y(x, w) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j \]
Sparse model

Model: \( y = Ax + n \), \( x \) is sparse

- \( y \): measurements, \( A \): dictionary
- \( n \): noise, \( x \): sparse weights
- Dictionary (A) – either from physical models or learned from data (dictionary learning)

\( N \ll M \)

\( N \times M \)

\( M \times 1 \) sparse signal

\( N \times 1 \) measurements

\( N \times 1 \) noise

\( k \)-sparsity
Sparse processing

• Linear regression (with sparsity constraints)
  – An underdetermined system of equations has many solutions
  – Utilizing $x$ is sparse it can often be solved
  – This depends on the structure of $A$ (RIP – Restricted Isometry Property)

• Various sparse algorithms
  – Convex optimization (Basis pursuit / LASSO / $L_1$ regularization)
  – Greedy search (Matching pursuit / OMP)
  – Bayesian analysis (Sparse Bayesian learning / SBL)

• Low-dimensional understanding of high-dimensional data sets

• Also referred to as compressive sensing (CS)
Different applications, but the same algorithm

Model: \( y = Ax + n \), \( x \) is sparse

\[
\begin{array}{c|c|c}
\text{Y} & \text{A} & \text{x} \\
\hline
\text{Frequency signal} & \text{DFT matrix} & \text{Time-signal} \\
\text{Compressed-Image} & \text{Random matrix} & \text{Pixel-image} \\
\text{Array signals} & \text{Beam weight} & \text{Source-location} \\
\text{Reflection sequence} & \text{Time delay} & \text{Layer-reflector} \\
\end{array}
\]
CS approach to geophysical data analysis

CS of Earthquakes
Yao, GRL 2011, PNAS 2013

Sequential CS
Mecklenbrauker, TSP 2013

CS beamforming
Xenaki, JASA 2014, 2015
Gerstoft JASA 2015

CS fathometer
Yardim, JASA 2014

CS Sound speed estimation
Bianco, JASA 2016

CS matched field
Gemba, JASA 2016
Sparse signals /compressive signals are important

• We don’t need to sample at the Nyquist rate

• Many signals are sparse, but are solved under non-sparse assumptions
  – Beamforming
  – Fourier transform
  – Layered structure

• Inverse methods are inherently sparse: We seek the simplest way to describe the data

• All this requires **new developments**
  - Mathematical theory
  - New algorithms (interior point solvers, convex optimization)
  - Signal processing
  - New applications/demonstrations
Sparse Recovery

- We try to find the sparsest solution which explains our noisy measurements.
- $L_0$-norm

$$1 \times L_0 = \sum_m 1_{x_m \neq 0}$$

- Here, the $L_0$-norm is a shorthand notation for counting the number of non-zero elements in $x$. 

Model: $y = Ax + n$, $x$ is sparse

\[ = \]

$N \times 1$ measurements $N \times M$

$M \times 1$ sparse signal

$1 \times L_0 = 3$
Sparse Recovery using $L_0$-norm

Underdetermined problem

$$y = Ax + n$$

$M < N$

Prior information

$x$: $K$-sparse, $K \ll N$

Not really a norm:

$$|a| = \sum_{n=1}^{N} 1_{x_n \neq 0} = K$$

There are only few sources with unknown locations and amplitudes

- $L_0$-norm solution involves exhaustive search
- Combinatorial complexity, not computationally feasible
L_p-norm

\[ \|x\|_p = \left( \sum_{m=1}^{M} |x_m|^p \right)^{1/p} \quad \text{for } p > 0 \]

\[ L_1 : \|x\|_1 = \sum_m |x_m| \]

\[ L_2 : \|x\|_2 = \left( \sum_m |x_m|^2 \right)^{1/2} \]

\[ L_\infty : p = \infty \]

- Classic choices for \( p \) are 1, 2, and \( \infty \).

- We will misuse notation and allow also \( p = 0 \).
$L_p$-norm (graphical representation)

$$\| x \|_p = \left( \sum_{m=1}^{M} |x_m|^p \right)^{1/p}$$
Solutions for sparse recovery

- Exhaustive search
  - \( L_0 \) regularization, not computationally feasible

- Convex optimization
  - \( L_1 \) regularization / Basis pursuit / LASSO

- Greedy search
  - Matching pursuit / Orthogonal matching pursuit (OMP)

- Bayesian analysis
  - Sparse Bayesian Learning (SBL)

- Regularized least squares
  - \( L_2 \) regularization, reference solution, not actually sparse
Regularized least squares

\[ \tilde{E}(w) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2 + \frac{\lambda}{2} \|w\|_2^2 \]

The squared weights penalty is mathematically compatible with the squared error function, giving a closed form for the optimal weights:

\[ w^* = (\lambda I + X^T X)^{-1} X^T t \]

A picture of the effect of the regularizer

- Slides 8/9, Lecture 4
- Regularized least squares solution
- Solution not sparse

The overall cost function is the sum of two parabolic bowls.
- The sum is also a parabolic bowl.
- The combined minimum lies on the line between the minimum of the squared error and the origin.
- The L2 regularizer just shrinks the weights.

\[ w_1, w_2 \neq 0 \]
Basis Pursuit / LASSO / $L_1$ regularization

- The $L_0$-norm minimization is not convex and requires combinatorial search making it computationally impractical.

- We make the problem convex by substituting the $L_1$-norm in place of the $L_0$-norm.

\[
\min_{x} \| x \|_1 \quad \text{subject to} \quad \| Ax - b \|_2 < \varepsilon \]

- This can also be formulated as

\[
\min_{x} \| Ax - y \|_2^2 + \mu \| x \|_1 \]

Regularizer \quad \text{convex opt.}

$X$ - Sparse

Sparsity is a fraction of $\mu$. 

CVX - MATLAB

\[
\begin{align*}
\text{Sparsity} & \quad \text{is fraction of} \quad \mu.
\end{align*}
\]
The unconstrained -LASSO- formulation

Constrained formulation of the $\ell_1$-norm minimization problem:

$$\hat{x}_{\ell_1}(\varepsilon) = \arg \min_{x \in \mathbb{C}^N} \|x\|_1 \text{ subject to } \|y - Ax\|_2 \leq \varepsilon$$

Unconstrained formulation in the form of least squares optimization with an $\ell_1$-norm regularizer:

$$\hat{x}_{\text{LASSO}}(\mu) = \arg \min_{x \in \mathbb{C}^N} \|y - Ax\|_2^2 + \mu \|x\|_1$$

For every $\varepsilon$ exists a $\mu$ so that the two formulations are equivalent.

Regularization parameter : $\mu$
Basis Pursuit / LASSO / $L_1$ regularization

• Why is it OK to substitute the $L_1$-norm for the $L_0$-norm?

• What are the conditions such that the two problems have the same solution?

$$\min_x \|x\|_1 \quad \text{subject to} \quad \|Ax - b\|_2 < \varepsilon$$

$$\min_x \|x\|_0 \quad \text{subject to} \quad \|Ax - b\|_2 < \varepsilon$$

• Restricted Isometry Property (RIP)

$$(1 - \delta_s)\|u\|_2 \leq \|A_S u\|_2 \leq (1 + \delta_s)\|u\|_2$$
Geometrical view of the lasso compared with a penalty on the squared weights.

L₂ regularization

L₁ regularization
Regularization parameter selection

The objective function of the LASSO problem:

$$\min \quad L(x, \mu) = \| y - Ax \|_2^2 + \mu \| x \|_1$$

- Regularization parameter: $\mu$
- Sparsity depends on $\mu$
  - $\mu$ large, $x = 0$
  - $\mu$ small, non-sparse

We can predict the jump in support

$$\| x \|_0 = k$$

$\mu \to \infty$$

$\| x \|_1^2$
- As regularization parameter $\mu$ is decreased, more and more weights become active.
- Thus $\mu$ controls sparsity of solutions.
Applications

- MEG/EEG/MRI source location (earthquake location)
- Channel equalization
- Compressive sampling (beyond Nyquist sampling)
- Compressive camera!
- Beamforming
- Fathometer
- Geoacoustic inversion
- Sequential estimation
Beamforming / DOA estimation

DOA estimation with sensor arrays

\[ y_m = \sum_{n} x_n e^{j \frac{2\pi}{\lambda} r_m \sin \theta_n} \]

\( m \in [1, \ldots, M] \): sensor

\( n \in [1, \ldots, N] \): look direction

\[ y = Ax \]

\[ \Theta : -90^\circ, 90^\circ \]

\[ y = [y_1, \ldots, y_M]^T, \quad x = [x_1, \ldots, x_N]^T \]

\[ A = [a_1, \ldots, a_N] \]

\[ a_n = \frac{1}{\sqrt{M}} \left[ e^{j \frac{2\pi}{\lambda} r_1 \sin \theta_n}, \ldots, e^{j \frac{2\pi}{\lambda} r_M \sin \theta_n} \right]^T \]

The DOA estimation is formulated as a linear problem
Direction of arrival estimation

Plane waves from a source/interferer impinging on an array/antenna

True DOA is sparse in the angle domain

$\Theta = \{0, \cdots, 0, \theta_1, 0, \cdots, 0, \theta_2, 0, \cdots, 0\}$

Amplitudes
Conventional beamforming

Plane wave weight vector $\mathbf{w}_i = [1, e^{-i \sin(\theta_i)}, \ldots, e^{-i(N-1) \sin(\theta_i)}]^T$

$$B(\theta) = |\mathbf{w}^H(\theta) \mathbf{b}|^2$$

ULA, half-wavelength spacing, $N = 20$ sensors, $\theta_1 = 20^\circ$, $\theta_2 = 30^\circ$,
Conventional beamforming

Equivalent to solving the $\ell_2$ problem with $A = [w_1, \cdots, w_M]$, $M > N$.

$$\min \|x\|_2^2 \text{ subject to } Ax = b, \|x\|_2 ^2 < \varepsilon$$

$A$ is an overcomplete dictionary of candidate DOA vectors. Columns span $-90^\circ$ to $90^\circ$ in steps of $1^\circ$ ($M = 181$).
\( \ell_1 \) minimization

In contrast \( \ell_1 \) minimization provides a sparse solution with exact recovery:

\[
\min \|x\|_1 \quad \text{subject to } \|Ax - b\|_2 \leq \varepsilon
\]

Columns of \( A \) span \(-90^\circ\) to \(90^\circ\) in steps of \(1^\circ\) \((M = 181)\).
Additional Resources

- A Mathematical Introduction to Compressive Sensing
  - Simon Foucart
  - Holger Rauhut

- Sparse and Redundant Representations
  - Michael Elad

- Compressed Sensing
  - Theory and Applications
  - Edited by Yonina C. Eldar and Gitta Kutyniok