AMATH 840: Advanced Numerical Methods for Computational and Data Science

Giang Tran
Department of Applied Mathematics, University of Waterloo

Winter 2024

Part 1: Sparse Optimization and Compressive Sensing

1.1: Sparse Solutions of Underdetermined Systems

Winter 2024

Outline

Solving Underdetermined Linear System $A\mathbf{w} = \mathbf{y}$

- Underdetermined system $A\mathbf{w} = \mathbf{y}$: Least Squares Solution
- Underdetermined system $A\mathbf{w} = \mathbf{y}$: Sparse Solutions

Introduction to Compressive Sensing

- Main Questions
- Compressive Sensing Models
- Applications, Motivations, and Extensions

Linear Models: $\mathbf{y} = A\mathbf{w}$

- Given $\mathbf{y} \in \mathbb{C}^m$ and a measurement matrix $A \in \mathbb{C}^{m \times N}$,
$$\text{find } \mathbf{w} \in \mathbb{C}^N : A\mathbf{w} = \mathbf{y}.$$
- The linear system may be inconsistent (no solutions), have one solution, or have infinitely many solutions.
- For the remaining of this section, we only study consistent linear systems.
    - The solution set is of the form
    $$\mathbf{w}^{\#} + \mathcal{W},$$
    where $A\mathbf{w}^{\#} = \mathbf{y}$ and $\mathcal{W}$ is an $(N - \text{rank}(A))$ dim. subspace of $\mathbb{C}^N$.
    - When $m \geq N$ and A is full column rank (i.e., $\text{rank}(A) = N$), then
    $$\dim \mathcal{W} = N - \text{rank}(A) = 0 \Rightarrow \mathcal{W} = \{\vec{0}\}.$$
    Therefore the consistent linear system has a unique solution.
    - For example, the Nyquist-Shannon sampling theorem says that the sampling rate of a continuous-time signal must be twice its highest frequency in order to ensure reconstruction.

Underdetermined system $A\mathbf{w} = \mathbf{y}, \quad (m < N)$

- If the consistent linear system is underdetermined, i.e., $m < N$, then the number of free variables is

$$N - \text{rank}(A) \geq N - \min\{m, N\} = N - m > 0.$$

  Therefore, the linear system has infinitely many solutions.

- In practice, we are only interested in finding a specific solution to the system $\Rightarrow$ Need additional information.

- Example 1: Given $A \in \mathbb{R}^{m \times N}, \mathbf{y} \in \mathbb{R}^m$ and $\text{rank}(A) = m$. Then $A\mathbf{x} = \mathbf{y}$ has at least one solution. Moreover, the solution with the smallest Euclidean norm is unique and is given by

$$\mathbf{w}_{ls} = A^T(AA^T)^{-1}\mathbf{y}.$$

Underdetermined system $A\mathbf{w} = \mathbf{y}$: Least Squares Solution

Proof.

- Suppose $\mathbf{w}_{ls}$ is a solution of $A\mathbf{x} = \mathbf{y}$ and $\mathbf{w}_{ls}$ has the smallest Euclidean norm among all solutions of the system $A\mathbf{x} = \mathbf{y}$.
- Then $\mathbf{w}_{ls}$ is a solution of the following optimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^N} \frac{1}{2}\|\mathbf{w}\|_2^2 \quad s.t. \quad A\mathbf{w} = \mathbf{y}.$$

- The corresponding Lagrangian is

$$L = \frac{1}{2}\|\mathbf{w}\|_2^2 + \lambda^T(\mathbf{y} - A\mathbf{w}), \quad \text{where} \quad \lambda \in \mathbb{R}^m.$$

- At the critical point $(\mathbf{w}_{ls}, \lambda_{ls})$, we have:

$$0 = \nabla_{\mathbf{w}_{ls}} L = \mathbf{w}_{ls} - A^T\lambda_{ls} \Rightarrow \mathbf{w}_{ls} = A^T\lambda_{ls},\ \mathbf{y} = A\mathbf{w}_{ls} = AA^T\lambda_{ls}$$

- Since $\text{rank}(AA^T) = \text{rank}(A) = m$ and $AA^T \in \mathbb{R}^{m\times m}$, (AA^T) is invertible.
- Therefore

$$\lambda_{ls} = (AA^T)^{-1}\mathbf{y} \quad \text{and} \quad \mathbf{w}_{ls} = A^T(AA^T)^{-1}\mathbf{y}.$$

□

Underdetermined system $A\mathbf{w} = \mathbf{y}$: Least Squares Solution

- Matlab code: $w = A\backslash y$.
- Python code:

```
import numpy as np
# Load A and y .... #
w_ls = np.linalg.lstsq(A, y, rcond=None)[0]
```

Underdetermined system $A\mathbf{w} = \mathbf{y}$: Least Squares Solution

Recall: Given $A \in \mathbb{R}^{m\times N}, \mathbf{y} \in \mathbb{R}^m$ and $\mathrm{rank}(A) = m$. Consider

$$\min_{\mathbf{w}\in\mathbb{R}^N} \frac{1}{2}\|\mathbf{w}\|_2^2 \quad s.t. \quad A\mathbf{w} = \mathbf{y}.$$

- Pros: A closed-form solution, $\mathbf{w}_{ls} = A^T(AA^T)^{-1}\mathbf{y}$.
- Cons:
  1. Least squares solutions likely overfit the data (See the codes).
  2. Least squares solutions are not robust to noisy measurements.
  3. In many applications, the solution with smallest Euclidean norm is not the expected solution. For example, reconstruct a one-dimensional discrete signal $f : \{1, \dots, N\} \to \mathbb{C}$ from a partial collection of its Fourier coefficients $\{\hat{f}(\xi_1), \dots, \hat{f}(\xi_m)\}$. Note that $m < N$.

Underdetermined system $A\mathbf{w} = \mathbf{y}$: Sparse Solution

- Another assumption: **w** is a sparse vector, i.e., most components of **w** are 0. Note that we don't know the locations of the nonzero entries. The problem can be recast as

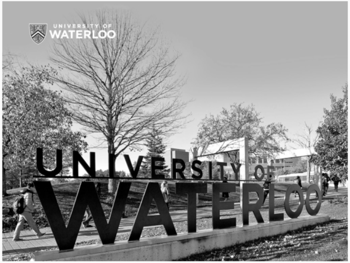
$$\min \|\mathbf{w}\|_0 \quad s.t. \quad A\mathbf{w} = \mathbf{y}.$$

- Does the sparsity assumption valid?
  - Related to simplicity, bet-on-sparsity principle, sparsity-of-effects principle, Pareto principle.
    - A system is usually dominated by main effects and low-order interactions.
    - Pareto principle: 80/20 rule or the law of the vital few.
  - Many real-world signals and images are compressible, i.e., well-approximated by sparse signals after an appropriate change of basis: MP3 signals, JPEG images,...
- In many applications, measurements are expensive or time consuming. [1]

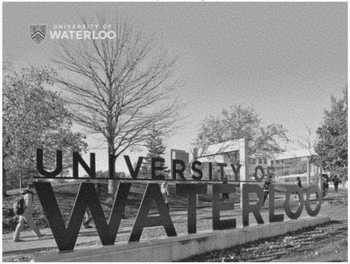
[1] See Section 1.2, "A Mathematical Introduction to Compressive Sensing", by Foucart and Rauhut.

Example: Sampling Theory

Original Gray Image



Compressed Image



Original Gray Image and Compressed Image (by keeping around 10% highest absolute values of the Fourier coefficients)

Outline

Compressive Sensing Problems

- Goal: Compress and Sense (acquire) data at the same time.
  - Acquire the compressed version of a signal directly via much fewer measured data than the signal length.
  - Reconstruct an s-sparse vector $\mathbf{w} \in \mathbb{C}^N$ from an underdetermined system $\mathbf{y} = A\mathbf{w} \in \mathbb{C}^m$, where $m \ll N$.
- Terminologies: Compressive sensing, compressed sensing, compressive sampling, sparse recovery.
- Definition: A vector $\mathbf{w} \in \mathbb{C}^N$ is called s-sparse if at most s of its entries are nonzero.
- Challenges: The locations of the non-zero entries of $\mathbf{w}$ is unknown $\longrightarrow$ Introduce the nonlinearity.

Compressive Sensing Problems

- Main questions:

  1. What matrices A are suitable?
     - Compressed sensing is not suitable for arbitrary matrices A.
     - Need to design a suitable linear measurement process.

  2. What is the minimal number of measurements? $\leftarrow$ Should depends on the compressed size, not on its uncompressed size!

  3. What are efficient (fast, stable, robust) reconstruction algorithms? Some popular ones are:
     - ℓ_0-algorithms: Orthogonal Matching Pursuit (OMP), Iterative Hard Thresholding (IHT), Hard Thresholding Pursuit (HTP)
     - ℓ_1-Optimization algorithms: FISTA, spgl1, Split Bregman/ADMM, SpaRSA, first-order primal-dual algorithm with linesearch.

Explain why compressed sensing is not suitable for arbitrary matrices A.

Compressive Sensing Problem: Models

- ℓ_0-minimization: NP-hard in general.

$$\min \|\mathbf{w}\|_0 \quad s.t. \quad A\mathbf{w} = \mathbf{y}.$$

- ℓ_1-minimization (convex relaxation of the ℓ_0-minimization):

$$\min \|\mathbf{w}\|_1 \quad s.t. \quad A\mathbf{w} = \mathbf{y} \quad \text{(Basis Pursuit)}.$$

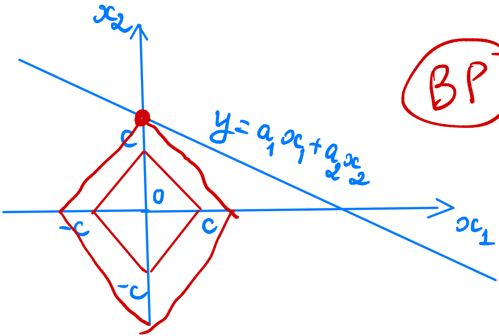
  Other models:

  - $\min \|\mathbf{w}\|_1 \quad s.t. \quad \|A\mathbf{w} - \mathbf{y}\|_2 \leq \eta,$
  - $\min \lambda\|\mathbf{w}\|_1 + \|A\mathbf{w} - \mathbf{y}\|_2^2,$
  - $\min \|A\mathbf{w} - \mathbf{y}\|_2^2 \quad s.t. \quad \|\mathbf{w}\|_1 \leq \tau.$

Why ℓ_1 for Sparsity?

- The ℓ_1-norm $\|\cdot\|_1$ is a convex function $\Rightarrow$ The ℓ_1-minimization problem can be solved by efficient algorithms from convex optimization.

- Ilustration of the ℓ_1-minimization induces sparse solutions:



BP

$n = 2$ $\quad m = 1$

$\min_{z \in \mathbb{R}^2} \|z\|_1$ s.t $y = Az$ $= a_1 z_1 + a_2 z_2 \in \mathbb{R}$

$\{\|z\|_1 = |z_1| + |z_2| = t,\ t \geq 0\} \cap \{z : y = Az\}$

Why ℓ_1 for sparsity

Theorem (Theorem 3.1.[2])

Let $A \in \mathbb{R}^{m\times N}$ be a measurement matrix with columns $\mathbf{a}_1, \ldots, \mathbf{a}_N$. Assuming the uniqueness of a minimizer $\mathbf{x}^\#$ of

$$\min_{\mathbf{z}\in\mathbb{R}^N} \|\mathbf{z}\|_1 \quad s.t. \quad A\mathbf{z} = \mathbf{y},$$

then the system $\{\mathbf{a}_j,\ j \in \mathrm{supp}(\mathbf{x}^\#)\}$ is linearly independent. In particular,

$$\|\mathbf{x}^\#\|_0 = \mathrm{card}(\mathrm{supp}(\mathbf{x}^\#)) \leq m.$$

[2] "A Mathematical Introduction to Compressive Sensing", by Foucart and Rauhut.

Compressive Sensing Problems

- Other problems:

4. Robustness: The output measurements are contaminated by noise. Find a spare vector $\mathbf{w}$ from $(A, \mathbf{y})$ such that

$$\mathbf{y} = A\mathbf{w} + \varepsilon, \quad \|\varepsilon\|_2 \leq \eta.$$

How is the solution affected by the noise?

5. Stability: $\mathbf{w}$ is not sparse, but is well-approximated by a sparse vector (compressibility).

Applications, Motivations, and Extensions

1. Applications:
    1.1 Magnetic Resonance Imaging
    1.2 Radar
2. Motivations:
    2.1 Sampling theory
    2.2 Sparse approximation
    2.3 Statistics and machine learning
3. Extensions:
    3.1 Low-rank recovery
    3.2 Matrix completion

Application: Magnetic Resonance Imaging [3]

- Goal: achieve high-resolution images based on few samples

- Observe: Angiogram images are sparse in the pixel representation; more complicated images have a sparse representation after transforming into a suitable domain such as wavelets (mathematically, $\mathbf{x} = \mathbf{W}\mathbf{x}'$ for some unitary matrix $\mathbf{W} \in \mathbb{C}^{N \times N}$ and some sparse vector $\mathbf{x}' \in \mathbb{C}^N$)

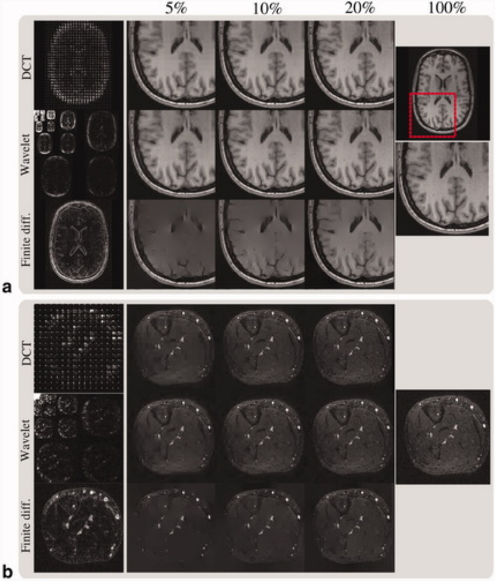
- Image reconstruction models: For example,

$$\min \|\psi \mathbf{x}\|_1 \quad \text{s.t.} \quad \|\mathcal{F}_u \mathbf{x} - \mathbf{y}\|_2 < \varepsilon,$$

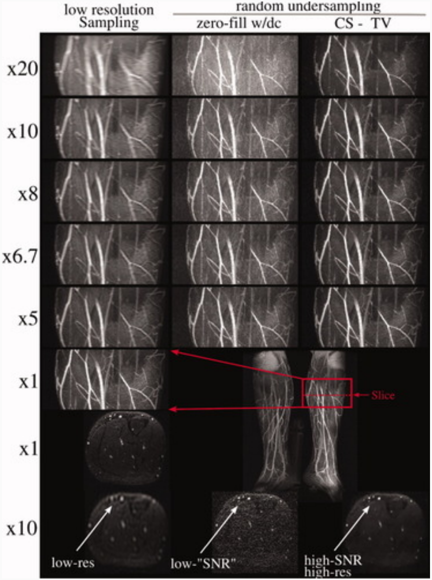
  where ψ is the linear operator that transforms from pixel representation into a sparse representation; $\mathcal{F}_u$ is a suitable undersampled Fourier transform, $\mathbf{x}$ is the reconstructed image, and $\mathbf{y}$ is the measured k-space data from the scanner.

- Compressive sensing results: Images with a sparse representation can be recovered from randomly undersampled k-space data, provided an appropriate nonlinear recovery scheme is used.

[3] Sparse MRI: The application of compressed sensing for rapid MR imaging, by M. Lustig, D. Donoho, and J. M. Pauly, `https://onlinelibrary.wiley.com/doi/full/10.1002/mrm.21391`



Transform-domain sparsity of images. (a) Axial T1 weighted brain image; (b) axial 3D contrast enhanced angiogram of the peripheral leg. The DCT, wavelet, and finite-differences transforms were calculated for all the images (Left column). The images were then reconstructed from a subset of 5, 10, and 20% of the largest transform coefficients. Source: Figure 3 from "Sparse MRI: The application of compressed sensing for rapid MR imaging", by M. Lustig, D. Donoho, and J. M. Pauly.



Contrast-enhanced 3D angiography reconstruction results as a function of acceleration. Left column: acceleration by LR. Note the diffused boundaries with acceleration. Middle column: ZF-w/dc reconstruction. Note the increase of apparent noise with acceleration. Right column: CS reconstruction with TV penalty from randomly undersampled k-space. Source: Figure 9 from "Sparse MRI: The application of compressed sensing for rapid MR imaging", by M. Lustig, D. Donoho, and J. M. Pauly.

Sparse Approximation and Compressive Sensing

- Consider a set $\{\mathbf{a}_1, \ldots, \mathbf{a}_N\} \subset \mathbb{C}^m$ s.t. $\text{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_N\} = \mathbb{C}^m$, called a dictionary. So $N \geq m$. For example, a union of several bases.
- Note that a dictionary may be linearly dependent. Indeed, redundancy may be desired when linearly independence is too restrictive.
- So, a representation $\mathbf{y} = \sum_{j=1}^{N} w_j \mathbf{a}_j$, where $\mathbf{w} = [w_1, \ldots, w_N]^T \in \mathbb{C}^N$, is not unique.
- Consider a representation with the smallest number of terms – a sparest representation.
- Mathematically, let $A \in \mathbb{C}^{m \times N}$ with columns $\mathbf{a}_1, \ldots, \mathbf{a}_N$ and $\mathbf{y} \in \mathbb{C}^m$. Solve

$$\min \|\mathbf{w}\|_0 \quad \text{s.t.} \quad A\mathbf{w} = \mathbf{y}.$$

Sparse Approximation and Compressive Sensing (cont'd)

Compressive Sensing	**Sparse Approximation**
• Free to design A with appropriate properties	• A is prescribed
• Estimate $\lVert \mathbf{x}_{original} - \mathbf{x}^{\#}_{reconstructed} \rVert$	• Estimate $\lVert \mathbf{y}_{given} - \mathbf{y}^{\#}_{\text{sparse expansion}} \rVert$

Statistics & Machine Learning and Compressive Sensing

- Statistical regression: Given an output data $\mathbf{y} \in \mathbb{R}^m$ and a data matrix $A \in \mathbb{R}^{m \times N}$, estimate $\mathbf{w} \in \mathbb{R}^N$ from

$$\mathbf{y} = A\mathbf{w} + \mathbf{e}.$$

  Here $\mathbf{e}$ is a random noise vector.

- In practice, the number N of parameters is often much larger than the number of observations.

- Selecting relevant explanatory variables (model selection problem is equivalent to find a sparse vector $\mathbf{w}$:

$$\min \|\mathbf{w}\|_1 \quad \text{s.t} \quad \|A\mathbf{w} - \mathbf{y}\|_2 \leq \eta.$$

  Other models:

$$\min \|A\mathbf{w} - \mathbf{y}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{w}\|_1 \leq \tau.$$

$$\min \frac{1}{2}\|A\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1.$$

$$\min \|\mathbf{w}\|_1 \quad \text{s.t} \quad \|A^*(A\mathbf{w} - \mathbf{y})\|_\infty \leq \lambda.$$

References

- A Mathematical Introduction to Compressive Sensing, by S. Foucart and H. Rauhut. Chapters 1 and 2.
- Statistical Learning with Sparsity, The Lasso and Generalizations, by T. Hastie, R. Tibshirani, and M. Wainwright. Chapter 1.
- Data-driven Science and Engineering, Machine Learning, Dynamical Systems, and Control, by S. L. Brunton and J. N. Kutz.
- Talks and lectures by T. Tao, S. Foucart, R. Willett, S. Brunton, H. Schaeffer,....

Exercises

1. Let $A \in \mathbb{R}^{m \times N}$. Prove that
$$\operatorname{rank}(A) = \operatorname{rank}(A^T) = \operatorname{rank}(A^T A) = \operatorname{rank}(AA^T).$$
2. Recall about pseudo inverse of a matrix, SVD.
3. Least squares problems (Prop. A.20.) Let $A \in \mathbb{C}^{m \times n}$ and $\mathbf{y} \in \mathbb{C}^m$. Define $\mathcal{M} := \{\mathbf{x} \in \mathbb{C}^n \quad \text{s.t. } \mathbf{x} \text{ is a minimizer of } \min \|A\mathbf{z} - \mathbf{y}\|_2\}$. Then the optimization problem
$$\min_{\mathbf{x} \in \mathcal{M}} \|x\|_2$$
has the unique solution $\mathbf{x}^\# = A^\dagger \mathbf{y}$.
4. Let $A \in \mathbb{C}^{m \times n}$ s.t. $\operatorname{rank}(A) = \min\{m, n\}$ (full rank) and let $\mathbf{y} \in \mathbb{C}^m$. Then
   4.1 If $m \geq n$, then the least squares problem
   $$\min_{\mathbf{x} \in \mathbb{C}^n} \|A\mathbf{x} - \mathbf{y}\|_2$$
   has the unique solution $\mathbf{x}^\# = A^\dagger \mathbf{y}$.
   4.2 If $n \geq m$, then the least squares problem
   $$\min_{\mathbf{x} \in \mathbb{C}^n} \|x\|_2 \quad \text{s.t.} \quad A\mathbf{x} = \mathbf{y}$$
   has the unique solution $\mathbf{x}^\# = A^\dagger \mathbf{y}$.