Lecture 11: Compressed Sensing

Xiaotie Deng

AIMS Lab
Department of Computer Science
Shanghai Jiaotong University

December 4, 2017

Overview

Big Data Reduction

- Design an algorithm that generates a given data, precisely.
  - Kolmogorov complexity.
- Law of large numbers:
  - Sampling theory in Statistics
- Sparse approximation:
  - Compression exploits sparsity in data to represents signals by few non-zero coefficients in a suitable basis, e.g., wavelet basis, and JPEG.
- Compressive sensing:
  - Data collection at reduced dimension.
  - Nyquist- Shannon sampling theorem for lower bound and J-L theorem for upper bound.
  - Data sparsity exploration.

Data Sparsity and Compression

- Finding a concise representation of a signal within acceptable distortion.
  - s-sparse data: $||x||_0 \leq s$ with $s << n$ nonzero coefficients, for a signal of length n.
  - to explore the computational efficiency further by making fewer measurement
- Key idea: to directly sense the data in a compressed form.
- Central Challenge: design of measurement schemes for applications to practical data models and acquisition systems

Difference from Classic Sampling

- Focus on measuring finite-dimensional vectors
- Acquire measurements in the form of inner products between the signal and more general test functions.
- Signal recovery is achieved using highly nonlinear methods.

Recovery Algorithms

- Nonlinear optimization recovers such signals from few measurements.
  - from a small set of measurements
  - provide performance guarantees for a variety of sparse recovery algorithms.
- Design low-dimensional signal models
- Accurately recover a high-dimensional signal
  - from a small set of measurements
  - provide performance guarantees for a variety of sparse recovery algorithms.

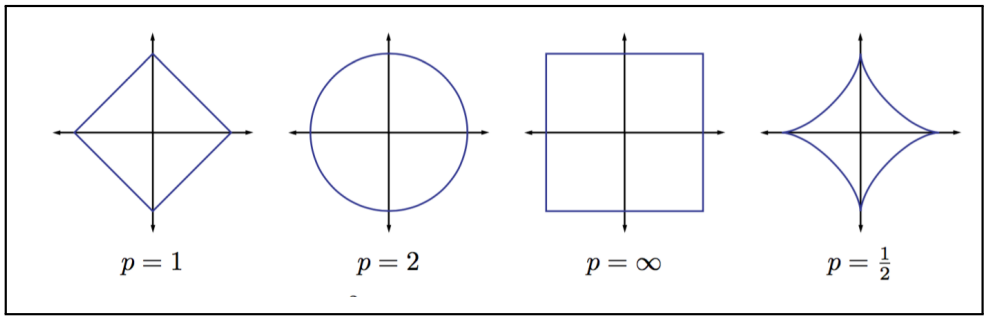
References

- E. Candès and T. Tao. Decoding by linear programming. IEEE Trans. Inform. Theory, 51(12):42034215, 2005.
- D. Donoho. Compressed sensing. IEEE Trans. Inform. Theory, 52(4):12891306, 2006.
- E.Candès.Compressive sampling.InProc.Int.Congress of Math., Madrid,Spain, Aug. 2006.
- R. Baraniuk. Compressive sensing. IEEE Signal Processing Mag., 24(4):118120, 124, 2007.

Modeling of Signal

Vector Space

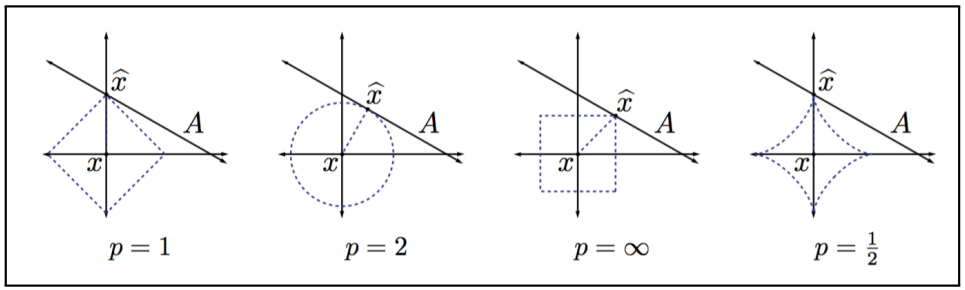
- Signal as a vector to represent the world's linear structure.
- Normed vectors $||x||_p = (\sum_i |x_i|^p)^{1/p}$
  - $||x||_0 = |support(x)|$, the number of none zero entries.
  - $||x||_\infty = \max_{all\ i} |x_i|$
- Unit sphere in L_1, L_2, L_∞ and $L_{1/2}$



- Inner product of x and z: $< x, z >= \sum_i x_i z_i$

Best Approximation

- Norm for the strength of a signal or significance of error.
- Goal of approximation: Find an approximation of the minimum error.
  - $\hat{x} = \min_{t \in A} ||t - x||_p$



- Obtained by growing the sphere radius in l_p-metric.

Bases

- Basis: $\{\phi_i\}_{i=1}^n$ if they are linearly independent, orthogonal if $(\phi_i, \phi_j) = \delta_{i,j}$ ($\Phi^T\Phi = I$).
- Any vector x can be uniquely represented by a basis: $x = \Phi c = \sum_{i=1}^n c_i\phi_i$, where $c_i = (x, \phi_i)$ for orthogonal basis.

Frames

- Frame: $\{\phi_i\}_{i=1}^n \subseteq R^d$ is a frame if $\exists 0 < A < B < \infty$, $\forall x: \ A||x||_2^2 \leq ||\Phi x||_2^2 \leq B||x||_2^2$.
  - $A^* = \sup\{A : \forall x \ A||x||_2^2 \leq ||\Phi x||_2^2\}$ and $B^* = \inf\{B : \forall x \ ||\Phi x||_2^2 \leq B||x||_2^2\}$ are called optimum frame bounds.
  - A-tight frame if $A^* = B^*$.
  - Parseval frame if $A = B = 1$.
  - Equal-norm if $\exists \lambda : \forall i \in \{1, 2, \cdots, n\} \ ||\phi_i||_2^2 = \lambda$
  - Unit norm if $\lambda = 1$.

Optimum Frame Bound

- Frame: $\Phi = (\phi_1, \phi_2, \cdots, \phi_n)$, a $d \times n$ matrix, is a frame if $\exists 0 < A < B < \infty$, $\forall x: \ A||x||_2^2 \leq ||\Phi x||_2^2 \leq B||x||_2^2$.
- It is representative because of its redundancy: $y = \Phi c$ may have many solutions of c.
- Optimum Frame Bound: The minimum and the maximum eigenvalues of $\Phi\Phi^T$.
  - $A^* = \sup\{A : \forall x \ A||x||_2^2 \leq ||\Phi x||_2^2\}$
  - $B^* = \inf\{B : \forall x \ ||\Phi x||_2^2 \leq B||x||_2^2\}$
- Canonical dual frame: $\tilde{\Phi} = \Phi^T(\Phi\Phi^T)^{-1}$. $\Phi\tilde{\Phi} = \tilde{\Phi}\Phi = I$
- $c_d = \arg\min_c\{||x - \hat{x}||_2 : c^T\Phi = \hat{x}\}$

MoorePenrose inverse

- $\Phi = (\phi_1, \phi_2, \cdots, \phi_n)$ is a $d \times n$ matrix, $d \leq n$.
- For Φ with all independent rows, $\Phi\Phi^T$ is full rank.
- Canonical dual frame: $\tilde{\Phi} = \Phi^T(\Phi\Phi^T)^{-1}$.
  - $\tilde{\Phi}$ is a right inverse (Moore-Penrose Inverse): $\Phi\tilde{\Phi} = I$.
  - Properties: $\tilde{\Phi}\Phi\tilde{\Phi} = \tilde{\Phi}$, $\Phi\tilde{\Phi}\Phi = \Phi$, $(\tilde{\Phi}\Phi)^T = \tilde{\Phi}\Phi$ and $(\Phi\tilde{\Phi})^T = \Phi\tilde{\Phi}$

Positive Eigenvalues of $\Phi\Phi^T$ and $\Phi^T\Phi$

- Both $\Phi\Phi^T$ and $\Phi^T\Phi$ are symmetric and positive semidefinite with all eigenvalues nonnegative.
- They have the same positive eigenvalues.
  - If x is the eigenvector of $\Phi^T\Phi$ wrt $\lambda > 0$: $(\Phi^T\Phi)x = \lambda x$; then $\Phi\Phi^T(\Phi x) = \lambda(\Phi x)$, implying Φx an eigenvector of $\Phi\Phi^T$ wrt λ.
  - Similarly, $(\Phi\Phi^T)y = \lambda y \implies \Phi^T\Phi(\Phi^T y) = \lambda(\Phi^T y)$ for $\lambda > 0$.
- Eigenvector wrt to zero eigenvalue:
  - None for $\Phi\Phi^T$ as it is full rank.
  - For $(\Phi^T\Phi)x = 0 \cdot x = 0$, $\Phi\Phi^T(\Phi x) = 0 \implies \Phi x = 0$ as $\Phi\Phi^T$ is of full rank.

Eigenvectors of of $\Phi\Phi^T$ and $\Phi^T\Phi$

- $\Phi\Phi^T$ has all eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_d$ positive, with corresponding orthogonal normal eigenvectors $y_1, y_2, \cdots, y_d$: of $\Phi\Phi^T y_i = \lambda_i y_i$.
- $\Phi^T\Phi$ has all eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_d, 0^{n-d}$ with corresponding orthogonal normal eigenvectors $x_1, x_2, \cdots, x_d; x_{d+1}, \cdots, x_n$: such that $\Phi^T\Phi x_i = \lambda_i x_i$.
- Properties
  - $\forall i \in \{1, 2, \cdots, d\}$: $x_i = \frac{1}{\sqrt{\lambda_i}}\Phi^T y_i$, $y_i = \frac{1}{\sqrt{\lambda_i}}\Phi x_i$.
  - Proof: Given $\Phi^T\Phi x_i = \lambda_i x_i$, set $y_i = \frac{1}{\sqrt{\lambda_i}}\Phi x_i$. $y_i^T y_i = \frac{1}{\lambda_i} x_i^T \Phi^T \Phi x_i = 1$, and $y_i^T y_j = 0$ for $i \neq j$.
- $(x_1, x_2, \cdots, x_n)^T (x_1, x_2, \cdots, x_n) = I \implies$ $I = (x_1, x_2, \cdots, x_n)(x_1, x_2, \cdots, x_n)^T = \sum_{i=1}^n x_i x_i^T$.

Singular Value Decomposition

- From $I = \sum_{i=1}^{n} x_i x_i^T$, $\Phi = \sum_{i=1}^{n} \Phi x_i x_i^T = \sum_{i=1}^{d} \Phi x_i x_i^T$
  - $||\Phi x_i||_2^2 = x_i^T \Phi^T \Phi x_i = 0$ for $i > d$ (since $\lambda_i(\Phi^T \Phi) = 0$.)
- $\Phi = \sum_{i=1}^{d} \Phi x_i x_i^T = \sum_{i=1}^{d} \sqrt{\lambda_i} y_i x_i^T$
- (right) Inverse $\Phi^G = \sum_{i=1}^{d} \frac{1}{\sqrt{\lambda_i}} x_i y_i^T$
- $\Phi\Phi^G = \sum_{i=1}^{d} \sum_{j=1}^{d} \sqrt{\lambda_i} \frac{1}{\sqrt{\lambda_j}} y_i x_i^T x_j y_j^T = \sum_{i=1}^{d} y_i y_i^T =$
  $(y_1, y_2, \cdots, y_d)(y_1, y_2, \cdots, y_d)^T = I$

Low-Dimensional Signal Models

Sparse models

- Use a linear combination of just a few elements from a known basis
- Non-linear approximation: k-sparse ($\Sigma_k = \{x : ||x||_0 \leq k\}$).
- Applications
  - compression, denoising, avoiding overfitting.
- image compression and image denoising by wavelet transform.
  - Recursively dividing image into its low- and high-frequency components
  - The lowest frequency components provide a coarse scale approximation of the image, while the higher frequency components fill in the detail and resolve edges.

Geometry of sparse signals and compressible signals

- Linear combination of two k-sparse signals may not be k-sparse any more.
- Measure of approximately sparse for compressible signals
  - $\sigma_k(x)_p = \min_{\hat{x} \in \Sigma_k} ||x - \hat{x}||_p$
- If $x = \Phi c$ the sorted sequence $|c_1| \geq |c_2| \cdots \geq \cdots \geq |c_n|$ obey power law decay if $|c_i| \leq C_1 i^{-q}$.

Finite unions of subspaces

- A finite union of subspaces model: the number of subspaces comprising the union is finite, and each subspace has finite dimensions: $\mathcal{U} = \cup_i U_i$
  - Structured sparse supports: This class consists of sparse vectors that meet additional restrictions on the support (i.e., the set of indices for the vectors nonzero entries): This corresponds to only certain subspaces Ui out of the $\binom{n}{k}$ subspaces present in Σ_k being allowed.
  - Sparse union of subspaces where each subspace $\mathcal{U}_i$ comprising the union is a direct sum of k low-dimensional subspaces $\mathcal{U}_i = \oplus_{j=1}^{k} \mathcal{A}_{i_j}$

Example: analog signal models

- Design new sensing systems for acquiring continuous-time, analog signals or images.
  - Represented continuos time discretely by a finite-length vector consisting of its Nyquist-rate samples: Xampling.
  - Another example of a signal class that can often be expressed as a union of subspaces is the class of signals having a finite rate of innovation
- Possibilities
  - finite unions of infinite dimensional spaces;
  - infinite unions of finite dimensional spaces;
  - infinite unions of infinite dimensional spaces.

Low-rank matrix models

- Using singular value decomposition.
- It can be written as $M = \sum_{k=1}^{r} \sigma_k u_k v_k^T$
  - where $u_k \in R^{n_1}, v_k \in R^{n_2}$
  - The matrix has $r(n_1 + n_2 - r)$ degree of freedom.

Manifold and Parametric models

- General class of low-dimensional signal models including:
  - a k-dimensional continuously-valued parameter θ identified carry relevant information about a signal.
  - the signal $f(\theta) \in \mathcal{R}^n$ changes as a continuous (typically nonlinear) function of these parameters
- $\mathcal{M} = \{f(\theta) : \theta \in \Theta\}$, Θ is a k-dimensional space.

Manifold and Parametric models

- Measurement $y = Ax$ where x is the signal and $A_{m \times n}$ represents dimension reduction.
  - How should we design the sensing matrix A to ensure that it preserves the information in the signal x?
  - How can we recover the original signal x from measurements y?
- Consider a number of desirable properties that we might wish A to have.
- Provide some important examples of matrix constructions that satisfy these properties.

Sensing Matrix

Null space conditions

- Null space $\mathcal{N}(A) = \{z : Az = 0\}$
- Spark of a given matrix A: the smallest number of columns of A that are linearly dependent.
- Theorem: For any vector $y \in R^m$, there exists at most one signal $x \in \Sigma_k$ such that $y = Ax$ iff $spark(A) > 2k$.
- Necessity: Proof (by contradiction). $spark(A) \leq 2k$, $\exists h \in \mathcal{N}(A) \cap \Sigma_{2k}$. We write $h = (x - x')$ where $x \neq x'; x, x' \in \Sigma_k$. Then $hx = hx' = y$ contradicting to assumption that three is at most one $x \in \Sigma_k$ such that $y = Ax$.
- Sufficiency: Suppose $x, x' \in Sigma_k$, $Ax = Ax'$. Then $h = x - x' \in \mathcal{N}(A) \cap \Sigma_{2k}$. As $spark(A) > 2k$ $Ah = 0$ implies $h = 0$. $Ax = y$ can have at most one solution x.

Approximately sparse signals

- $\Lambda \subset \{1, 2, \cdots, n\}$, define $\Lambda^c = \{1, 2, \cdots, n\} - \Lambda$.
- $\forall i \in \Lambda : x_\Lambda(i) = x(i); \forall j \in \Lambda^c : x_\Lambda(j) = 0.$ and A_Λ sets columns of A indexed in Λ^c to zero.
- NSP: Define a matrix A satisfies the null space property of **order** k iff there is some $C > 0$ such that $||h_\Lambda||_2 \leq C \frac{||h_{\Lambda^c}||_1}{\sqrt{k}}$ for all $h \in \mathcal{N}(A)$ and for all Λ such that $|\Lambda| \leq k$.
- Property: the only k-sparse vector for an NSP matrix is 0.
  - Proof: Let A satisfies NSP. Consider $h \in \Sigma_k \cap \mathcal{N}(A)$. Taking $\Lambda = \{i : h(i) \neq 0\}$, $h_{\Lambda^c} = 0$. By NSP, it implies $h_\Lambda = 0$.

Instance-optimal recovery

- Let $\Delta : R^m \rightarrow R^n$ be the recovery method.
- Focusing on guarantee of the form $||\Delta(Ax) - x||_2 \leq C \frac{\sigma_k(x)_1}{\sqrt{k}}$ for all x, where $\sigma_k(x)_1$ is defined as $\min_{\hat{x} \in \Sigma_k} ||x - \hat{x}||_1$.
- It guarantees the recovery of all k-sparse signals.
- Theorem: Let $A : R^n \rightarrow R^m$ be a sensing matrix and $\Delta : R^m \rightarrow R^n$ an arbitrary recovery algorithm. Then (A, Δ) satisfies $||h_\Lambda||_2 \leq C \frac{||h_{\Lambda^c}||_1}{\sqrt{k}}$ then A satisfies NSP of order 2k.

Proof of Theorem

- Let $h \in \mathcal{N}(A)$. Let Λ be the indices corresponding to the $2k$ largest entries of h.
- Split $\Lambda = \Lambda_0 \cup \Lambda_1$: $|\Lambda_0| = |\Lambda_1| = k$.
- Set $x = h_{\Lambda_1} + h_{\Lambda^c}$ and $x' = -h_{\Lambda_0}$ so that $h = x - x'$.
- As $x' \in \Sigma_k$, $x' = \Delta(Ax')$ from above.
- $h \in \mathcal{N}(A)$ implies $Ah = 0$ and $Ax = Ax' \rightarrow x' = \Delta(Ax)$.
- $||h_\Lambda||_2 \leq ||h||_2 = ||x - x'||_2 = ||x - \Delta(Ax)||_2 \leq C\frac{\sigma_k(x)_1}{\sqrt{k}} = \sqrt{2}C\frac{||h_{\Lambda^c}||_1}{\sqrt{2k}}$.

The restricted isometry property (RIP)

- A satisfies *RIP* of order k iff $\exists \delta_k \in (0,1)$ such that $\forall x \in \Sigma_k : (1-\delta_k)||x||_2^2 \leq ||Ax||_2^2 \leq (1+\delta_k)||x||_2^2$
- NOTE this is equivalent to the existence of $\alpha, \beta > 0$ such that $\forall x \in \Sigma_k : \alpha||x||_2^2 \leq ||Ax||_2^2 \leq \beta||x||_2^2$ by multiplying A by $\sqrt{2/(\alpha+\beta)}$ to obtain $\delta = \frac{\beta-\alpha}{\beta+\alpha}$.
- Corollary: Suppose that A satisfies the RIP of order k with constant δ_k. Let γ be a positive integer. Then A satisfies the RIP of order $k' = \gamma\lfloor\frac{k}{2}\rfloor$ withconstant $\delta_{k'} < \gamma\delta_k$.

RIP and Stability

- RIP is sufficient for a variety of algorithms to be able to successfully recover a sparse signal from noisy measurements
- Stability: Let (A, Δ) be the sensing and recovery pair. It is C-stable if $\forall x \in \Sigma_k$ and any $e \in R^m$, we have $||\Delta(Ax + e) - x||_2 \leq C||e||_2$, which implies
  - Theorem: $\frac{1}{C}||x||_2 \leq ||Ax||_2$
- Proof: Take $x, z \in \Sigma_k$ and Define $e_x = \frac{A(z-x)}{2}$ and $e_z = \frac{x-z}{2}$. Then $Ax + e_x = Az + e_z = \frac{A(x+z)}{2}$.
- Let $\hat{x} = \Delta(Ax + e_x) = \Delta(Az + e_z$.
- $||x - z||_2 = ||x - \hat{x}||_2 + ||\hat{x} - z||_2 \leq C||e_x||_2 + C||e_z||_2 = C||Ax - Az||_2$, which holds for all $x, z \in \Sigma_{2k}$. Therefore, claim holds.

Measurement Bound for RIP

- Let A be an $m \times n$ matrix satisfy RIP of order 2k with constant $\delta \in (0, \frac{1}{2})$. Then
  - $m \geq Ck \log \binom{n}{k}$

  where $C = \frac{1}{2log(\sqrt{24}+1)} \sim 0.28$.
- Johnson-Lindenstrauss lemma is closely related to the RIP.

RIP implies NSP

- Let A be an $m \times n$ matrix satisfy RIP of order 2k with constant $\delta < \sqrt{2} - 1$. Then
- Then A satisfies the NSP of order 2k with constant $C = \frac{\sqrt{2}\delta_{2k}}{1-(1+\sqrt{2})\delta_{2k}}$.

Coherence

- Coherence of a matrix A is defined as $\mu(A) = \max_{1 \le i < j \le n} \frac{<a_i, a_j>}{||a_i||_2 ||a_j||_2}$
- Property: $\mu(A) \in [\sqrt{\frac{n-m}{m(n-1)}}, 1]$
- It is easier to compute than spark, NSP, RIP which needs to computer $\binom{n}{k}$ submatrices.

Coherence, Spark, RIP and NSP

- $\forall A : spark(A) \geq 1 + \frac{1}{\mu(A)}$.
- If $k < \frac{1}{2}(1 + \frac{1}{\mu(A)})$ then for each measurement vector $y \in R^m$, there exists at more one $x \in \Sigma_k$ such that $y = Ax$.
- If A has unit norm columns and coherence $\mu = \mu(A)$, then A satisfies RIP of order k with $\delta = (k-1)\mu$ for all $k < \frac{1}{\mu}$.

Sensing Matrix Construction

- We now construct A with the defined properties.
- Random matrices A of size $m \times n$ whose entries are independent and identically distributed (i.i.d.) with continuous distributions have $spark(A) = m + 1$ with probability one.
- It will satisfy the RIP with high probability if the entries are chosen according to a Gaussian, Bernoulli, or more generally any sub-gaussian distribution.
- These random constructions provide matrices satisfying the NSP.
- When the distribution used has zero mean and finite variance, then in the asymptotic regime (as m and n grow) the coherence converges to $\mu(A) = \sqrt{2 \log n)/m}$.

Assignments IV

- In the lecture we designed a left pseudoinverse of matrix $A_{m,n}$ for $m < n$. Prove or disprove A also has a left pseudoinverse.
- What is the coherence $\mu(A)$ of the identity matrix? Provide a tight bound of $\mu(A)$?
- Do the same for spark, RIP and NSP for the class of $2 \times n$ matrices.
- Find the worst case ratio matrix of $\mu(A)$ and spark(A).