Introduction to Topics in Machine Learning

Namrata Vaswani

Department of Electrical and Computer Engineering
Iowa State University

- Compressed Sensing / Sparse Recovery: Given $y := Ax$ recover x from y when y is shorter than x. Use sparsity of x.
- Low-rank Matrix Completion: Given a subset of entries of a low-rank matrix M, complete the matrix
  - given $y = \mathcal{P}_\Omega(M)$, find M. Ω: set of indices of the observed entries
- Matrix Sensing: given a set of n linear functions of M, find M using the fact that M is low-rank
  - given $y = \mathcal{A}(M)$ where $\mathcal{A}(.)$ is a linear operator, find M. This can be written as $y_i =< A_i, M >$ where $< A, B >= trace(A'B)$ is the usual inner product.
- Robust PCA: given $Y := X + L$, find X and L
  - $L =$ unknown low rank matrix.
  - $X =$ sparse matrix (corresponds to outliers)
- Phase retrieval: compute vector x from $y := |Ax|^2$. Here $|.|$ means element-wise magnitude of the vector. More specifically $y_i = |A^i x|^2$ (here A^i is the i-th row of A).

- the term phase retrieval comes from Fourier imaging where A is the DFT matrix; but now it's used more generally for any matrix A

- Ranking and individualized ranking estimation

Applications

- CS: projection imaging - MRI, CT, single-pixel camera, radar, ...
- MC: recommendation system design, e.g., Netflix problem
- Matrix sensing: one special case is phase retrieval. Notice that we can rewrite $y_i = A^i x x' A^{i'} =< xx', A^i A^{i'} >$
- RPCA: recommendation system design in the presence of outliers, Video analytics, Survey data analysis,
- Phase retrieval: astronomy, X-ray crystallography,...

Non-convex Problems: Alternating Minimization and Gradient Descent I

Alternating Min

- Goal: compute $\min_{x,y} f(x,y)$ when $f(.)$ is non-convex
- Clearly $\min_{x,y} f(x,y) = \min_x(\min_y f(x,y))$ but of course in most cases, RHS is also hard to compute.
- Consider the class of problems where the min is easy when one variable is fixed, i.e., $\min_y f(x_0, y)$ is easy for a given x_0 and $\min_x f(x, y_0)$ is easy for a given y_0.
  - A common solution: Alt-Min
  - Start with an initial guess x_0.
  - Compute $y_1 \in \arg\min_y f(x_0, y)$
  - Compute $x_1 \in \arg\min_y f(x, y_1)$
  - Repeat above until a stopping criterion is met.
- Guarantees? Till very recently none. Recent work:

Non-convex Problems: Alternating Minimization and Gradient Descent II

- If initialized carefully, Alt-Min gets to within a small error of the true solution in a finite number of iterations. Possible to bound this number also.
- A common approach to initialization: "spectral method" - compute the top eigenvector of an appropriately defined matrix
- Guarantees exist for Matrix Completion and for Phase Retrieval

Gradient descent based approaches for non-convex problems

- With a suitable initialization, it is possible to get a guarantee
- Truncated gradient descent idea of "truncated Wirtinger flow" paper: the gradient turns out to be a weighted average of certain vectors; discard those weights that are too large and compute a truncated gradient estimate
-

Old slides on sparse recovery

The sparse recovery / compressed sensing problem

- Given $y := Ax$ where A is a fat matrix, find x.
  - underdetermined system, without any other info, has infinite solutions
- Key applications where this occurs: Computed Tomography (CT) or MRI
  - CT: acquire radon transform of cross-section of interest
  - typical set up: obtain line integrals of the cross-section along a set of parallel lines at a given angle, and repeated for a number of angles from 0 to π), common set up: 22 angles, 256 parallel lines per angle
  - by Fourier slice theorem, can use radon transform to compute the DFT along radial lines in the 2D-DFT plane
  - Projection MRI is similar, directly acquire DFT samples along radial lines
  - parallel lines is most common type of CT, other geometries also used.
- Given 22x256 data points of 2D-DFT of the image, need to compute the 256x256 image

Limitation of zero-filling

- A traditional solution: zero filling + I-DFT
  - set the unknown DFT coeff's to zero, take I-DFT
  - not good: leads to spatial aliasing
- Zero-filling is the minimum energy (2-norm) solution, i.e. it solves $\min_x \|x\|_2 \ s.t. \ y = Ax$. Reason
  - clearly, min energy solution in DFT domain is to set all unknown coefficients to zero, i.e. zero-fill
  - (energy in signal) = (energy in DFT)*2π, so min energy solution in DFT domain is also the min energy solution
- The min energy solution will not be sparse because 2-norm is not sparsity promoting
  - In fact it will not be sparse in any other ortho basis either because $\|x\|_2 = \|\Phi x\|_2$ for any orthonormal Φ. Thus min energy solution is also min energy solution in Φ basis and thus is not sparse in Φ basis either
- But most natural images, including medical images, are approximately sparse (or are sparse in some basis)

Sparsity in natural signals/images

- Most natural images, including medical images, are approximately sparse (or are sparse in some basis)
  - e.g. angiograms are sparse
  - brain images are well-approx by piecewise constant functions (gradient is sparse): sparse in TV norm
  - brain, cardiac, larynx images are approx. piecewise smooth: wavelet sparse
- Sparsity is what lossy data compression relies on: JPEG-2000 uses wavelet sparsity, JPEG uses DCT sparsity
- But first acquire all the data, then compress (throw away data)
- In MRI or CT, we are just acquiring less data to begin with - can we still achieve exact/accurate reconstruction?

Use sparsity as a regularizer

- Min energy solution $\min_x \|x\|_2 \; s.t. \; y = Ax$ is not sparse, but is easy to compute $\hat{x} = A'(AA')^{-1}y$
- Can we try to find the min sparsity solution, i.e. find $\min_x \|x\|_0 \; s.t. \; y = Ax$
- Claim: If true signal, x_0, is exactly S-sparse, this will have a unique solution that is EXACTLY equal to x_0 if $spark(A) > 2S$
  - $spark(A) =$ smallest number of columns of A that are linearly dependent.
  - in other words, any set of (spark-1) columns are always linearly independent
- proof in class
- Even when x is approx-sparse this will give a good solution
- But finding the solution requires a combinatorial search: $O(\sum_{k=1}^{S} \binom{m}{k}) = O(m^S)$

Practical solutions [Chen,Donoho'95] [Mallat,Zhang'93]

- Basis Pursuit: replace ℓ_0 norm by ℓ_1 norm: closest norm to ℓ_0 that is convex

$$\min_x \|x\|_1 \ s.t. \ y = Ax$$

- Greedy algorithms: Matching Pursuit, Orthogonal MP
- Key idea: all these methods "work" if columns of A are sufficiently "incoherent"
- "work": give exact reconstruction for exactly sparse signals and zero noise, give small error recon for approx. sparse (compressible) signals or noisy measurements

Compressive Sensing

- name: instead of capturing entire signal/image and then compressing, can we just acquire less data?
- i.e. can we compressively sense?
- MRI (or CT): data acquired one line of Fourier projections at a time (or random transform samples at one angle at a time)
- if need less data: faster scan time
- new technologies that use CS idea:
  - single-pixel camera,
  - A-to-D: take random samples in time: works when signal is Fourier sparse
  - imaging by random convolution
  - decoding "sparse" channel transmission errors.
- Main contribution of CS: theoretical results

General form of Compressive Sensing

- Assume that an N-length signal, z, is S-sparse in the basis Φ, i.e. $z = \Phi x$ and x is S-sparse.
- We sense

$$y := \Psi z = \underbrace{\Psi\Phi} Ax$$

- It is assumed that Ψ is "incoherent w.r.t. Φ"
  - or that $A := \Psi\Phi$ is "incoherent"
- Find x, and hence $z = \Phi x$, by solving

$$\min_x \|x\|_1 \; s.t. \; y = Ax$$

- A random Gaussian matrix, Ψ, is "incoherent" w.h.p for S-sparse signals if it contains $O(S \log N)$ rows
- And it is also incoherent w.r.t. any orthogonal basis, Φ w.h.p. This is because if Ψ is r-G, then $\Psi\Phi$ is also r-G (ϕ any orthonormal matrix).
- Same property for random Bernoulli.

Quantifying "incoherence"

- Rows of A need to be "dense", i.e. need to be computing a "global transform" of x.
- Mutual coherence parameter, $\mu := \max_{i \neq j} |A_i' A_j| / \|A_i\|_2 \|A_j\|_2$
- spark(A) = smallest number of columns of A that are linearly dependent.
- Or, any set of $(spark(A) - 1)$ columns of A are always linearly independent.
- RIP, ROP
- many newer approaches...

Quantifying "incoherence": RIP

- A $K \times N$ matrix, A satisfies the S-Restricted Isometry Property if constant δ_S defined below is positive.
- Let A_T, $T \subset \{1, 2, \dots N\}$ be the sub-matrix obtained by extracting the columns of A corresponding to the indices in T. Then δ_S is the smallest real number s.t.

$$(1 - \delta_S)\|c\|^2 \leq \|A_T c\|^2 \leq (1 + \delta_S)\|c\|^2$$

  for all subsets $T \subset \{1, 2, \dots N\}$ of size $|T| \leq S$ and for all $c \in \mathbb{R}^{|T|}$.
  - In other words, every set of S or less columns of A has singular values b/w $\sqrt{1 \pm \delta_S}$
  - $\Leftrightarrow$ every set of S or less columns of A approximately orthogonal
  - $\Leftrightarrow$ A is approximately orthogonal for any S-sparse vector, c.

Examples of RIP

- If A is a random Gaussian, random Bernoulli, or Partial Fourier matrix with about $O(S \log N)$ rows, it will satisfy RIP(S) w.h.p.
- Partial Fourier * Wavelet: somewhat "incoherent"

Use for spectral estimation and comparison with MUSIC

- Given a periodic signal with period N that is a sparse sum of S sinusoids, i.e.

$$x[n] = \sum_k X[k] e^{j2\pi kn/N}$$

where the DFT vector, X, is a $2S$-sparse vector.

- In other words, x[n] does not contain sinusoids at arbitrary frequencies (as allowed by MUSIC), but only contains harmonics of $2\pi/N$ and the fundamental period N is known.
- In matrix form, $x = F^* X$ where F is the DFT matrix and $F^{-1} = F^*$.

- Suppose we only receive samples of $x[n]$ at random times, i.e. we receive $y = Hx$ where H is an "undersampling matrix" (exactly one 1 in each row and at most one 1 in each column)
- With random time samples it is not possible to compute covariance of $\underline{x}[n] := [x[n], x[n-1], \ldots x[n-M]]'$, so cannot use MUSIC or the other standard spectral estimation methods.
- But can use CS. We are given $y = HF^* X$ and we know X is sparse. Also, $A := HF^*$ is the conjugate of the partial Fourier matrix and thus satisfies RIP w.h.p.
- If have $O(S \log N)$ random samples, we can find X exactly by solving

$$\min_X \|X\|_1 \; s.t. \; y = HF^* X$$

Quantifying "incoherence": ROP

- θ_{S_1,S_2}: measures the angle b/w subspaces spanned by A_{T_1}, A_{T_2} for disjoint sets, T_1, T_2 of sizes less than/equal to S_1, S_2 respectively
- $\theta_{S1,S2}$ is the smallest real number such that

$$|c1' A'_{T1} A_{T2} c2| < \theta_{S1,S2} \, \|c1\| \, \|c2\|$$

  for all $c1, c2$ and all sets $T1$ with $|T1| \leq S1$ and all sets $T2$ with $|T2| \leq S2$
- In other words

$$\theta_{S1,S2} = \min_{T1,T2:|T1|\leq S1,|T2|\leq S2} \min_{c1,c2} \frac{|c1' A'_{T1} A_{T2} c2|}{\|c1\| \, \|c2\|}$$

- Can show that δ_S is non-decreasing in S, θ is non-decreasing in $S1, S2$
- Also $\theta_{S1,S2} \leq \delta_{S1+S2}$
- Also, $\|A_{T_1}{}' A_{T_2}\| \leq \theta_{|T_1|,|T_2|}$

Theoretical Results

- If x is S-sparse, $y = Ax$, and if $\delta_S + \theta_{S,2S} < 1$, then basis pursuit exactly recovers x
- If x is S-sparse, $y = Ax + w$ with $\|w\|_2 \leq \epsilon$, and $\delta_{2S} < (\sqrt{2} - 1)$, then solution of basis-pursuit-noisy, $\hat{x}$ satisfies

$$\|x - \hat{x}\| \leq C_1(\delta_{2S})\epsilon$$

- basis-pursuit-noisy:

$$\min_x \|x\|_1 \ s.t. \ \|y - Ax\|_2 \leq \epsilon$$

MP and OMP

Applications

DSP applications

- Fourier sparse signals
  - Random sample in time
  - Random demodulator + integrator + uniform sample with low rate A-to-D
- N length signal that is sparse in any given basis Φ
  - Circularly convolve with an N-tap all-pass filter with random phase
  - Random sample in time or use random demodulator architecture

Compressibility: one definition

Papers to Read

- Decoding by Linear Programming (CS without noise, sparse signals)
- Dantzig Selector (CS with noise)
- Near Optimal Signal Recovery (CS for compressible signals)
- Applications of interest for DSP
  - Beyond Nyquist:... Tropp et al
  - Sparse MRI: ... Lustig et al
  - Single pixel camera: Rice, Baranuik's group
  - Compressive sampling by random convolution : Romberg

Sparse Recon. with Partial Support Knowledge

- Modified-CS (our group's work)
- Weighted ℓ_1
- von-Borries et al

Treating Outliers as Sparse Vectors

- Dense Error Correction via ell-1 minimization
- “Robust” PCA
- Recursive “Robust” PCA (our group’s work)