CS838 Topics In Optimization
Convex Geometry in High-Dimensional Data Analysis

Ben Recht
Spring 2010
Logistics

- Class Tu-Th 1-2:15PM
- Office Hours CS4387, Tuesday 2:15-3:30

- Course webpage:  [http://pages.cs.wisc.edu/~brecht/cs838.html](http://pages.cs.wisc.edu/~brecht/cs838.html)
- Readings will be posted here.

- **Scribing.** All students are required to scribe notes for at least one lecture. LaTeX template will be provided.

- **Project.** All students are required to prepare a 20-30 minute presentation on the themes of this course. This can be a literature review or an application of the course’s techniques to your research.
Recommender Systems
Netflix

• Rate some movies...

• Get some recommendations:
Netflix Prize

• One million big ones!
Netflix Prize

- One million big ones!

- Given 100 million ratings on a scale of 1 to 5, predict 3 million ratings to highest accuracy

- 17770 total movies x 480189 total users
- Over 8 billion total ratings

- How to fill in the blanks?
Abstract Setup: Matrix Completion

- How do you fill in the missing data?

\[ X = \]

\[ X_{ij} \text{ known for black cells} \]
\[ X_{ij} \text{ unknown for white cells} \]

*Rows index movies*
*Columns index users*

\[ X = L R^* \]

- \( k \times n \) entries
- \( k \times r \) entries
- \( r \times n \) entries
- \( r(k+n) \) entries
Netflix Prize - Dimensions

\[ X = L R^* \]

- \( k = \) Number of movies = \( 2 \times 10^4 \)
- \( n = \) Number of users = \( 5 \times 10^5 \)
- \( m = \) Number of Given Ratings = \( 10^8 \)
- \( kn \approx 10^{10} \)

For \( r < 200 \), \( r(k+n) < 10^8 \)
Matrix Rank

- The rank of \( X \) is...
  - the dimension of the span of the rows
  - the dimension of the span of the columns
  - the smallest number \( r \) such that there exists an \( k \times r \) matrix \( L \) and an \( n \times r \) matrix \( R \) with \( X = LR^* \)
Constraints involving the rank of the Hankel Operator, Matrix, or Singular Values
Affine Rank Minimization

**PROBLEM:** Find the matrix of lowest rank that satisfies/approximates the underdetermined linear system

\[ A(X) = b \quad A : \mathbb{R}^{k \times n} \rightarrow \mathbb{R}^m \]

minimize \( \text{rank}(X) \)
subject to \( A(X) = b \)

**NP-HARD:**
- Reduce to finding solutions to polynomial systems
- Hard to approximate
- Exact algorithms are awful
Heuristic: Gradient Descent

\[
F(L, R) = \sum_{i=1}^{k} \sum_{k=1}^{r} L_{ik}^2 + \sum_{j=1}^{n} \sum_{k=1}^{r} R_{jk}^2 + \lambda \sum_{i,j} \left( \sum_{k} L_{ik} R_{jk} - X_{ij} \right)^2
\]

- Just run gradient descent to minimize \( F \)
- \( \lambda \) determines tradeoff between satisfying constraints and the size of the factors
Gradient descent on low-rank parameterization

Mixture of hundreds of models, including gradient descent
Complex Systems

Predictions

Structure

Rank

Smoothness

Sparsity
Modeling Simplicity: Strategy

• Find a “natural” convex heuristic

\[ tx_1 + (1 - t)x_2 \]

• Use probabilistic analysis to prove the heuristic succeeds

• Provide efficient algorithms for solving the heuristic
Topics

- Sparsity
- Rank
- Smoothness

Themes

- Random Projections Preserve Geometry (encoding)
- Atomic Norms Recover Geometry (decoding)
Parsimonious Models

- Search for best linear combination of fewest atoms
- "rank" = fewest atoms needed to describe the model

- "natural" heuristic is the atomic norm:

\[
\|x\|_A \equiv \inf \left\{ \sum_{k=1}^{r} |w_k| : x = \sum_{k=1}^{r} w_k \alpha_k \right\}
\]
Mining for Biomarkers

- $n_{\text{patients}} \ll n_{\text{peaks}}$
- If very few are needed for diagnosis, search for a sparse set of markers
- $l_1$, LASSO, etc.
Topic 1: Cardinality/Sparsity

- Vector $x$ has cardinality $s$ if it has at most $s$ nonzeros. ($x$ is $s$-sparse)

$$x = \sum_{k=1}^{s} w_k e_{i_k}$$

- Atoms are a discrete set of orthogonal points
- Typical Atoms:
  - standard basis
  - Fourier basis
  - Wavelet basis
Cardinality Minimization

- **PROBLEM:** Find the vector of lowest cardinality that satisfies/approximates the underdetermined linear system

\[ Ax = b \quad A : \mathbb{R}^n \to \mathbb{R}^m \]

- **NP-HARD:**
  - Reduce to EXACT-COVER [Natarajan 1995]
  - Hard to approximate
  - Known exact algorithms require enumeration
Proposed Heuristic

**Cardinality Minimization:**
\[
\begin{align*}
\text{minimize} & \quad \text{card}(x) \\
\text{subject to} & \quad Ax = b
\end{align*}
\]

**Convex Relaxation:**
\[
\begin{align*}
\text{minimize} & \quad \|x\|_1 = \sum_{i=1}^{D} |x_i| \\
\text{subject to} & \quad Ax = b
\end{align*}
\]

- Long history (back to geophysics in the 70s)
- Flurry of recent work characterizing success of this heuristic: Candès, Donoho, Romberg, Tao, Tropp, etc., etc...
- “Compressed Sensing”
Compressed Sensing

- Model: most of the energy is at low frequencies
- Basis for JPG compression
- Use the fact that the image is sparse in DCT/wavelet basis to reduce number of measurements required for signal acquisition.
- *decode using $l_1$ minimization*
Why $l_1$ norm?

$\|x\|_1$

$\text{card}(x)$
- 1-sparse vectors of Euclidean norm 1

- Convex hull is the unit ball of the $l_1$ norm

\[ \{ x : \| x \|_1 \leq 1 \} \]

\[ \| x \|_1 = \sum_{i=1}^{n} | x_i | \]
minimize \( \|x\|_1 = \sum_{i=1}^{n} |x_i| \)
subject to \( Ax = b \)
**Integer Programming**

- Integer solutions: all components of $x$ are $\pm 1$

- Convex hull is the unit ball of the $l_\infty$ norm

\[
\{ x : \| x \|_\infty \leq 1 \}
\]

\[
\| x \|_\infty = \max_i |x_i|
\]
minimize $\|x\|_\infty = \max_i |x_i|$
subject to $Ax = b$
Cardinality/Sparsity

• How many samples are required to reconstruct sparse vectors?
• Relationship to coding theory
• When can we guarantee the l1 heuristic works?
• What are efficient ways to compute minimum l1 norm solutions?
Topic 2: (Matrix) Rank

- Matrix $X$ has rank $r$ if it has at most $r$ nonzero singular values.

$$X = \sum_{j=1}^{r} \sigma_j u_j v_j^* = \sum_{j=1}^{r} \sigma_j A_j$$

- Atoms are the set of all rank one matrices
- Not a discrete set
Singular Value Decomposition (SVD)

- If $X$ is a matrix of size $k \times n$ ($k \leq m$) then there matrices $U$ ($k \times k$) and $V$ ($n \times k$) such that

\[
X = U \Sigma V^*
\]

\[
U^*U = I_m \quad V^*V = I_m
\]

- A diagonal matrix, $\sigma_1 \geq \ldots \geq \sigma_k \geq 0$

- $\sigma_i^2$ is an eigenvalue of $XX^*$. $U$ are eigenvectors of $XX^*$.

- **Fact:** If $X$ has rank $r$, then $X$ has only $r$ non-zero singular values.
SVD = Filter Bank

• Multiply a vector $z$ by $X = U \Sigma V^*$
Collaborative Filterings

- $Z$ is a linear combination of eigenusers, $v_1, \ldots, v_k$.
- $u_1, \ldots, u_k$ are the eigenratings.
Which Algorithm?

**Affine Rank Minimization:**
\[
\text{minimize} \quad \text{rank}(X) \\
\text{subject to} \quad A(X) = b
\]

**Convex Relaxation:**
\[
\text{minimize} \quad \|X\|_* = \sum_{i=1}^{k} \sigma_i(X) \\
\text{subject to} \quad A(X) = b
\]

- Proposed by Fazel (2002).
- Nuclear norm is the “numerical rank” in numerical analysis.
- The “trace heuristic” from controls if $X$ is p.s.d.
Why nuclear norm?

- Just as $l_1$ norm induces sparsity, nuclear norm induces low rank
- Nuclear norm of diagonal matrix = $l_1$ norm of diagonal
• 2x2 matrices
• plotted in 3d

\[
\begin{bmatrix}
  x & y \\
  y & z
\end{bmatrix}
\]

rank 1

\[x^2 + z^2 + 2y^2 = 1\]

Convex hull:

\[\{X : \|X\|_* \leq 1\}\]
• 2x2 matrices
• plotted in 3d

\[ \left\| \begin{bmatrix} x & 0 \\ 0 & z \end{bmatrix} \right\|_1 \leq 1 \]

• Projection onto x-z plane is $l_1$ ball
- 2x2 matrices
- plotted in 3d

\[ \| \begin{bmatrix} x & y \\ y & z \end{bmatrix} \| \leq 1 \]

- Not polyhedral...

So how do we compute it? And when does it work?
Computationally: Gradient Descent!

\[ \mathcal{F}(L, R) = \sum_{i=1}^{k} \sum_{j=1}^{r} L_{ij}^2 + \sum_{i=1}^{n} \sum_{j=1}^{r} R_{ij}^2 + \lambda \| A(LR^*) - b \|^2 \]

- “Method of multipliers”
- Schedule for \( \lambda \) controls the noise in the data
- Same global minimum as nuclear norm
Topic 2: Rank

- How many samples are required to reconstruct low-rank matrices?
- Fast algorithms for SVD as compressed sensing
- When can we guarantee the nuclear norm heuristic works?
- What are efficient ways to compute minimum nuclear norm solutions?
Goal: Find a class $\mathcal{F}$ which is easy to search over, but can approximate complex behavior.

Typically a list of example inputs

Which space of functions? dictated by application

Learning Functions

$$\min_{f \in \mathcal{F}} \text{fitness}(f, \text{data})$$

Error bound:
$$\text{error} < O(n^{-\frac{s}{d}})$$

Number of examples

Curse of dimensionality

Blessing of smoothness
Topic 3: Approximation

- Try to write a function as a sum of (non-orthogonal) bases:

\[
f(x) \approx \sum_{k=1}^{n} c_k \phi_k(x; \theta_k)
\]

- Atoms are sets of basis functions
- Not a discrete set, infinite dimensional space.
• **Goal:** Find a class $\mathcal{F}$ which is easy to search over, but can approximate complex behavior.

• **Solution:** Approximate $f(x)$ by $f_n(x) = \sum_{k=1}^{n} c_k \phi_k(x; \theta_k)$.
• Approximate \( f(x) \) by \( f_n(x) = \sum_{k=1}^{n} c_k \phi_k(x; \theta_k) \)

For large class of \( f \), sampling \( \theta_k \) i.i.d. and optimizing \( c_k \) yields

\[
\|f - f_n\| = O\left(\frac{1}{\sqrt{n}}\right)
\]

Analysis via convex hull norm where the atoms are \( \phi(x; \theta) \)

Radial Basis Functions

\[
\phi(x; \omega, b) = \cos(\omega'x + b)
\]

\[
\omega \sim \mathcal{N}(0, 1)
\]

\[
b \sim \text{unif}[-\pi, \pi]
\]

\[
k(x, y) = \exp(-\gamma\|x - y\|^2)
\]
% Approximates Gaussian Process regression
% with Gaussian kernel of variance gamma
% lambda: regularization parameter
% dataset: X is dxN, y is 1xN
% test: xtest is dx1
% D: dimensionality of random feature

% training
w = randn(D, size(X,1));
b = 2*pi*rand(D,1);
Z = cos(sqrt(gamma)*w*X + repmat(b,1,size(X,2)));
alpha = (lambda*eye(size(X,2)+Z*Z')\(Z*y);

% testing
ztest = alpha(:)'*cos( sqrt(gamma)*w*xtest(:) + ...
    + repmat(b,1,size(X,2)) );
Topic 3: Approximation

- How many bases are required to approximate complicated behavior?
- What are efficient ways to fit functions in infinite dimensional function spaces?
- What are fast ways to fit functions when we are overwhelmed by data?