Large-Scale Optimization

L12. COMPRESSED SENSING II
Today

Topics:

• Compressed sensing: sparse representation of data
• Concentration of measure in high dimensions
• Matrix completion
Compressed Sensing

An inverse problem of dimensionality reduction: can we reconstruct the original signal from observations?

\[ y = A \in \mathbb{R}^{k \times n} \quad x \in \mathbb{R}^n \]

Observations

Sensing matrix

Original signal

s-sparse

(Figure adapted from R. Baraniuk's talk slides)
Sparsity by Change of Representation

Yale Face Dataset

- [http://vision.ucsd.edu/content/yale-face-database](http://vision.ucsd.edu/content/yale-face-database)
- 15 people, each with 11 different conditions
- Dim: $243 \times 320 = 77760$
- “Normal” faces:

![Normal faces from the Yale Face Database](image-url)
Change of Representation

Representation based on pixels

\[ \begin{align*} &\quad = \quad + \quad + \quad + \quad + \quad + \quad + \quad + \quad + \quad + \quad + \quad + \quad \ldots \end{align*} \]

Representation based on “singular faces”

\[ \begin{align*} &\quad = \quad -228.9 \quad -27.3 \quad + \quad + \quad + \quad + \quad + \quad + \quad + \quad + \quad + \quad + \quad + \quad + \quad + \quad \ldots \end{align*} \]

TU Dortmund, Dr. Sangkyun Lee
Singular Faces

$\text{SVD} = \begin{bmatrix} U & D & V^T \end{bmatrix}$

an original image | singular faces | combination coeffs

$U^T U = UU^T = I$

$V^T V = VV^T = I$

Reduced SVD

$\begin{bmatrix} U & D & V^T \end{bmatrix} = \begin{bmatrix} U \end{bmatrix}$
Sparse Representation

Images can be represented as a sparse vector, when represented in a proper basis.
Universality of CS

Random measurements can be used for signals sparse in any basis.
DCT: Discrete Cosine Transform

8x8 (JPEG) Basis of 2D-DCT

Change of representation (basis)

\[ \begin{array}{c}
1 = 26 \text{ [light grey]} + 12 \text{ [dark grey]} + \ldots \\
\end{array} \]

An 8 x 8 patch of

How many coefficients do we have for each 8 x 8 patch?
Sparsifying in Cosine Domain

Original → DCT

In each 8x8 block, make small coeffs=0

Density = 1.42%

“Sparsified” ← IDCT
Compressed Sensing

\[ y = A \alpha \]

\[ n = 76 \times 100 \quad k \geq s \log(n/s) \]

\[
\min_{x \in \mathbb{R}^n} \| y - Ax \|_2^2 + \lambda \| x \|_1
\]

\[ A: \text{random Bernoulli} \]

ISTA, \# iter = 1000

Solution had 100% correct support

Density = 1.42%
Random Projection Perspective

\[ y = A \in \mathbb{R}^{k \times n} \cdot x \in \mathbb{R}^n \]

unknown signal

query

\[ y_i = \langle x, A_i \rangle \text{ if } \|A_i\| = 1, \]

answer to the query
Knowing $x$ by Projection Queries

Q: what kind of queries would work?

Q. What kind of queries be the best, to minimize the number of queries?

Q: how many queries do we need?

$$k \approx s \log(n/s) \ll n$$
Curse of High Dimensionality

Sampling:

• Volume to be sampled increases exponentially with extra dimensions

Machine Learning:

• Algorithms require a very large number of extra training examples to compensate an increase in dimensionality

...
Generating Good Queries

How difficult would it be to generate a good query matrix $A$?

How about a random matrix $A$?
Gaussian Random Vectors

$\mathcal{N}(0, .2)$
Bernoulli Random Vectors

\[
\{+1, -1\}^n
\]
“Blessing” of High Dimensionality

Concentration of measure phenomenon

• V. Milman
• Lipschitz functions on random variables $X$ in dim $n$ following distributions on a sphere concentrate on their expectation
• Extended to Gaussian, etc.
Generalized Gradient Descent for Constrained Cases

Consider a constrained convex optimization problem,

\[
\min_{x \in C} F(x)
\]

where \( F \) is convex function and \( C \) is a convex set

This is equivalent to solve:

\[
\min_{x \in \mathbb{R}^n} F(x) + I_C(x)
\]

\[
I_C(x) = \begin{cases} 
0 & \text{if } x \in C \\
\infty & \text{if } x \notin C
\end{cases}
\]
Projected Gradient Descent

\[
\min_{x \in \mathbb{R}^n} F(x) + I_C(x) \quad \quad \quad I_C(x) = \begin{cases} 
0 & \text{if } x \in C \\
\infty & \text{if } x \notin C 
\end{cases}
\]

\[
\text{prox}_{\alpha_k I_C}(x) = \arg \min_z \left\{ \frac{1}{2} \|x - z\|^2 + \alpha_k I_C(z) \right\}
\]

\[
= \arg \min_{z \in C} \frac{1}{2} \|x - z\|^2 = P_C(x)
\]

\[
x_{k+1} = \text{prox}_{\alpha_k I_C}(x_k - \alpha_k \nabla F(x_k)) = P_C(x_k - \alpha_k \nabla F(x_k))
\]

A gradient descent is taken and projected onto the feasible set
This is called projected gradient descent
Matrix Completion

Original m x n matrix A

- We observe only the entries $A_{ij}, (i, j) \in \Omega$
- Sparse in its spectrum = low rank

Matrix completion: fill-in the remaining entries by

$$
\min_{X \in \mathbb{R}^{m \times n}} \frac{1}{2} \sum_{(i,j) \in \Omega} (A_{ij} - X_{ij})^2 + \lambda \|X\|_*
$$

$$
\|X\|_* = \sum_{i=1}^{r} \sigma_i(X)
$$
Define $[P_{\Omega}]_{ij} = \begin{cases} X_{ij} & (i, j) \in \Omega \\ 0 & (i, j) \notin \Omega \end{cases}$

$$\min_{X \in \mathbb{R}^{m \times n}} \frac{1}{2} \sum_{(i, j) \in \Omega} (A_{ij} - X_{ij})^2 + \lambda \|X\|_*$$

$$f(X) = \frac{1}{2} \|P_{\Omega}(A) - P_{\Omega}(X)\|_F^2 + \lambda \|X\|_*$$

$$\nabla F(X) = -(P_{\Omega}(A) - P_{\Omega}(X))$$

$$\text{prox}_{\alpha_k \psi} = \arg \min_{Z \in \mathbb{R}^{m \times n}} \frac{1}{2} \|X - Z\|_F^2 + \alpha_k \lambda \|Z\|_*$$
Matrix Soft-Thresholding Operator

$$\text{prox}_{\alpha_k} \psi (X) = S_{\lambda \alpha_k} (X)$$

$$S_{\lambda} (X) := U \Sigma \lambda V^T$$

where $$X = U \Sigma V^T$$ and

$$[\Sigma_{\lambda}]_{ii} = \max\{\Sigma_{ii} - \lambda, 0\}$$

A Singular Value Thresholding Algorithm for Matrix Completion, Cai, Candes, and Shen, SIAM J OPT, 2010
Proximal Gradient Descent for Matrix Completion

\[ \nabla F(X) = -(P_\Omega(A) - P_\Omega(X)) \]

\[ \text{prox}_{\alpha_k \psi}(X) = S_{\lambda \alpha_k}(X) \]

\[ X_{k+1} = S_{\lambda \alpha_k}(X_k - \alpha_k \nabla F(X_k)) \]
\[ = S_{\lambda \alpha_k}(X_k + \alpha_k (P_\Omega(A) - P_\Omega(X))) \]

\( \nabla F \) is Lipschitz continuous with \( L = 1 \), so we can choose \( \alpha_k = 1 \),

\[ X_{k+1} = S_{\lambda}(P_\Omega(A) + P^\perp_\Omega(X)) \]
\[ P^\perp_\Omega(X) = X - P_\Omega(X) \]

This is the soft-impute algorithm

- Mazumder, Hastie, and Tibshirani, Spectral regularization for learning large incomplete matrices, 2011
When Matrix Completion Works?

Essentially, MC will work when the observed entries of $A$ give good information about unobserved entries.

$$A = U \Sigma V^T$$

\[
\begin{align*}
A_{rj} &= U_r \cdot \Sigma V_{j.}^T \\
A_{ij} &= U_i \cdot \Sigma V_{j.}^T \\
A_{ik} &= U_i \cdot \Sigma V_{k.}^T
\end{align*}
\]
Matrix Completion: Incoherence

A) Let $P_U$ (respectively, $P_V$) be the orthogonal projection onto the singular vectors $u_1, \ldots, u_r$ (respectively, $v_1, \ldots, v_r$). For all pairs $(a, a') \in [n_1] \times [n_1]$ and $(b, b') \in [n_2] \times [n_2]$.

\[
\left| \langle e_a, P_U e_{a'} \rangle - \frac{r}{n_1} 1_{a=a'} \right| \leq \mu \frac{\sqrt{r}}{n_1} \\
\left| \langle e_b, P_V e_{b'} \rangle - \frac{r}{n_2} 1_{b=b'} \right| \leq \mu \frac{\sqrt{r}}{n_2}.
\]

B) Let $E$ be the “sign matrix” defined by

\[ E = \sum_{k \in [r]} u_k v_k^*. \]

For all $(a, b) \in [n_1] \times [n_2]$,

\[ |E_{ab}| \leq \mu \frac{\sqrt{r}}{\sqrt{n_1 n_2}}. \]

Matrix completion with noise, Candes & Plan, Proc. of IEEE, 2010
Recovery Theorem

Theorem 2 [14]: Let \( M \in \mathbb{R}^{n_1 \times n_2} \) be a fixed rank-\( r \) matrix with strong incoherence parameter \( \mu \), and set \( n := \max(n_1, n_2) \). Suppose we observe \( m \) entries of \( M \) with locations sampled uniformly at random. Then there is a numerical constant \( C \) such that if

\[
m \geq C\mu^2 nr \log^6 n
\]  

(II.8)

then \( M \) is the unique solution to (II.4) with probability at least \( 1 - n^{-3} \).

Matrix completion with noise, Candes & Plan, Proc. of IEEE, 2010
Compressed sensing:

- Talk slides by Richard Baraniuk,
- Talk slides by Mark Davenport,