EE 381V: Large Scale Optimization **Fall 2012**

Lecture 15 — March 5

Lecturer: Caramanis & Sanghavi *Scribe: Gezheng Wen*

15.1 Large Scale Linear Algebra

15.1.1 Direct vs Iterative Methods

- **Direct Methods**: Direct methods are widely used for matrix inversion, matrix factorization.
  - High cost to perform $\sim O(n^3)$;
  - Typically do not exploit special matrix structure (e.g., sparsity)
- **Indirect Methods**: Interative Algorithms
  - Often approximate to the true covariance;
  - Often inexpensive at each step(involve at most matrix-vector product)

  Note: This naturally exploits sparsity. Typical examples include power iteration and Lanczos algorithm.

15.1.2 Lanczos Review

Recall the key ideas of Lanczos algorithm are as follows: Given $Q_k = [q_1, q_2, ...q_k]$ where q_i are orthonormal vectors, $M_k \triangleq \lambda_1(Q_k^T A Q_k) = \max_{0 \neq y \in \mathbb{R}^{k \times 1}} \frac{(Q_k y)^T A (Q_k y)}{y^T y}$ approximates to $\lambda_1(A)$.

Given a vector $q \in \mathbb{R}^n$ and $A \in \mathbb{S}^n$, the Krylov subspace $\mathcal{K}(A, q, k)$ is given by: $\mathcal{K}(A, q, k) = \text{span}\left\{q, Aq, Aq^2, \ldots Aq^{k-1}\right\}$.
The Krylov matrix $K(A, q, k)$ is given by: $K(A, q, k) = \left[q \ Aq \ Aq^2 \ \ldots Aq^{k-1}\right]$.

One property of the Krylov space is $span\{q_1, q_2, ..., q_k\} = span\{q_1, Aq_1, ..., A^{k-1}q_1\}$.

Here is the summary of Lanczos iterations:

- The algorithm produces orthonormal basis for the Krylov space $\mathcal{K}(A, q, k)$ iteratively.
- From the Kaniel-Pagie Theory, we know: $\lambda_1(Q_k^T A Q_k) \approx \lambda_1(A)$, $\lambda_k(Q_k^T A Q_k) \approx \lambda_n(A)$.
- $Q_k^T A Q_k = \mathrm{T}_k$, where T_k is a tri-diagnoal and symmetric matrix.

All these nice properties of Lanczos iterations lead to the following consequences:

- Q_k and T_k are very easy to compute iteratively.
- After the Lanczos iteration, the eigenvalues of T_k are also easy to compute

In homework, we will prove:

1. It is possible to implement Lanczos iterations with a single matrix-vector multiplication.
2. If $A \in \mathbf{R}^{n\times n}$ is a s-sparse per row, then each Lanczos iteration requires about $O(n \cdot s)$.

15.1.3 Solving $Ax = b$ by Lanczos Iterations

There are two classes of fundamental problems in Linear Algebra

1. Finding the eigenvalues/eigenvectors of a matrix, or singular values/singular vectors of a matrix
2. Solving $\min \|Ax - b\|_2$, or $Ax = b$.

For the second class, if we let $\Phi(x) = \frac{1}{2}x^T Ax - x^T b$, where A is positive semidefinite matrix, then it is easy to see that $\nabla\Phi(x) = Ax - b \Rightarrow x^* = A^{-1}b$. Thus, solving $\min \|Ax - b\|_2$ and solving $Ax = b$ is essentially the same problem.

Now let us find out how we may use iterations to solve this problem.

$$\begin{aligned} Q_k &= [q_1, q_2, ..., q_k] & (15.1)\\ x_k &= x_0 + Q_k y_k & (15.2)\\ (Q_k^T A Q_k) y_k &= Q_k^T (b - Ax_0) & (15.3) \end{aligned}$$

To make this iteration algorithm efficient and easy to compute, we need solve three problems:

1. find an easy way to find q_{k+1}
2. find an easy solution to 15.3
3. find a Fast way to update x_k

Let us exploit Lanczos Alogrithm by recalling the Theorem: After k steps of Lanczos, we have $AQ_k = Q_k T_k + r_k e_k^T$, where $\mathrm{T}_k = \begin{bmatrix} \alpha_1 & \beta_1 & & & 0 \\ \beta_1 & \ddots & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{k-2} & \ddots & \beta_{k-1} \\ 0 & & & \beta_{k-1} & \alpha_k \end{bmatrix}$ and $Q_k = [q_1 \cdots q_k]$.

Problem 1: Find q_{k+1}

Use Lanczos iteration

Problem 2: Solve $(Q_k^T A Q_k)y_k = Q_k^T(b - Ax_0)$

Based on the theorem, we can solve 15.3 by solving $\mathrm{T}_k y_k = Q_k^T(b - Ax_0)$. Since Q_k is generated visa Lanczos iterations, T_k is symmetric and tri-diagonal. Thus, T_k can be factorized as:

$$\mathrm{T}_k = L_k D_k L_k^T, \tag{15.4}$$

where $L_k = \begin{bmatrix} 1 & & & & 0 \\ \mu_1 & \ddots & & & \\ & \ddots & \ddots & & \\ & & \mu_{k-2} & \ddots & \\ 0 & & & \mu_{k-1} & 1 \end{bmatrix}$, and $D_k = \begin{bmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_n \end{bmatrix}$.

By comparing the entries of the matrices, we can find the entries of L_k and D_k by iterations, and all L_k and D_k are easy to obtain from L_{k-1} and D_{k-1}.

$$\begin{aligned} d_1 &= \alpha_1, && (15.5) \\ \mu_{i-1} &= \frac{b_{i-1}}{d_{i-1}}, i = 2, 3, ..., k && (15.6) \\ d_i &= \alpha_i - \beta_{i-1}\mu_{i-1}, i = 2, 3, ..., k && (15.7) \end{aligned}$$

Problem 3: Update $x_k = x_0 + Q_k y_k$

Then we need to figure out the fast way to update x_k. Recall that each Lanczos iteration needs one matrix-vector product.

- **Question:** Can we avoid all the above matrix-vector products?
- **Answer:** Yes, it is possible to do again with only one matrix-vvector iteration per step.

One of the algorithms to accomplish the task is called *Conjugate Gradient Method.*
Define: $C_k \in \mathbf{R}^{n\times k}$, $p_k \in \mathbf{R}^k$ via

$$\begin{aligned} C_k L_k^T &= Q_k && (15.8) \\ L_k D_k L_k^T &= Q_k^T(b - Ax_0) = Q_k^T r_0, && (15.9) \end{aligned}$$

where $r(\hat{x}) = b - A\hat{x}, r_0 = b - Ax_0$. Setting $r_0 = b - Ax_0$, we can define the update of x_k via C_k and p_k because:

$$\begin{aligned} x_k &= x_0 + Q_k y_k \\ &= x_0 + Q_k(Q_k A Q_k^T)^{-1} * Q_k^T(b - Ax_0) \\ &= x_0 + Q_k \mathrm{T}_k^{-1} Q_k^T r_0 \\ &= x_0 + Q_k(L_k D_k L_k^T)^{-1} Q_k^T r_0 \\ &= x_0 + C_k p_k \end{aligned} \tag{15.10}$$

Now write $C_k = [C_{k-1}\, c_k]$, where c_k is the k^{th} column of C_k. From $C_k L_k^T = Q_k$, we can find c_k:

$$c_k = q_k - \mu_{k-1} e_{k-1} \tag{15.11}$$

Now $p_k \in \mathbf{R}^k \Rightarrow p_k = (\rho_1, \rho_2, ...\rho_k)^T$.

Recall p_k is defined by: $L_k D_k L_k^T = Q_k^T r_0$. This is equivalent to:

$$\begin{bmatrix} & L_{k-1}D_{k-1} & & 0 \\ 0\cdots & 0 & \mu_{k-1}d_{k-1} & d_k \end{bmatrix} \begin{bmatrix} \rho 1 \\ \rho 2 \\ \vdots \\ \rho_{k-1} \\ \rho_k \end{bmatrix} = \begin{bmatrix} q_1^T r_0 \\ q_2^T r_0 \\ \vdots \\ q_{k-1}^T r_0 \\ q_k^T r_0 \end{bmatrix}$$

Thus, the update of x_k can be expressed as:

$$\begin{aligned} x_k &= x_0 + C_k p_k \\ &= x_0 + [C_{k-1}\, c_k] \begin{bmatrix} p_{k-1} \\ \rho_k \end{bmatrix} \\ &= (x_0 + C_{k-1} p_{k-1}) + c_k p_k \\ &= x_{k-1} + c_k p_k \end{aligned} \tag{15.12}$$

Now we can update x via vector-scalar multiplication.

Since $L_{k-1} D_{k-1} p_{k-1} = Q_{K-1}^T r_0$, we have $\rho_k = (q_k^T r_0 - \frac{\mu_{k-1} d_{k-1} \rho_{k-1}}{d_k})$.

$$\begin{aligned} x_k &= x_0 + Q_k y_k = x_0 + C_k p_k \\ &= x_0 + C_{k-1} p_{k-1} + \rho_k c_k = x_{k-1} + \rho_k c_k \end{aligned} \tag{15.13}$$

In summary:

1. Lanczos:

$$\begin{aligned} T_{k-1} &\rightarrow T_k \ (\alpha_{k-1}\beta_{k-1} \rightarrow \alpha_k \beta_k) \\ q_{k-1} &\rightarrow q_k \end{aligned}$$

2. LDL^T factorizatoin of T_k gives:

$$\begin{aligned} \mu_{k-1} &= \frac{\beta_{k-1}}{d_{k-1}} \\ d_k &= \alpha_k - \beta k - 1\mu_{k-1} \end{aligned}$$

3. x_k update

$$\begin{aligned} c_k &= q_k - \mu_{k-1}C_{k-1} \\ \rho_k &= \frac{(q_k^T r_0 - \mu_{k-1}d_{k-1}\rho_{k-1})}{d_k} \\ x_k &= x_{k-1} + \rho_k c_k \end{aligned}$$

15.2 Spiked Covariance

An important problem in multivariate statistical analysis is the estimation of the covariance matrix. An interesting question to ask here would be how many samples are needed to estimate the covariance.

More precisely, if $X_i \sim \mu, i.i.d, X_i \in \mathbb{R}^p$, then the empirical covariance Σ can be calculated as $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$, where n is the number of samples. Note here that the expectation of $X_i X_i^T$ equals to the actual covariance (i.e., $\mathbb{E}[X_i(X_I)^T] = \Sigma$). In addition, from the standard Law of Large Numbers, we know $\hat{\Sigma} \sim \Sigma$ as $n \to \infty$.

Now let us consider this problem in high-dimensional settings. There are multiple possible scenarios when the number of dimension p and the number of samples s increase. Three typical examples of the relations are shown in (Figure 15.2). For this lecture, we are particularly interested in the cases that $\frac{p}{n} \to c$ (i.e., the number of samples increases in the approximately same rate as the dimensionality of the sample data points increases). We want to know what the empirical covariance $\hat{\Sigma}$ looks like.
Let us consider two cases:

- **Case 1**: The samples only contain noise. That means $\Sigma = I$
- **Case 2**: The samples contain both signal and noise. That means $\Sigma = I + \sigma w^T$

15.2.1 Case 1: noise only

Assume $n \to \infty$, $p \to \infty$, and $\frac{p}{n} \to c$, where p is the dimensionality of the samples,n is the number of samples and c is a constant. What do eigenvalues of the covariance matrix look like?
If the noise is Gaussian with zero means and identity covariance matrix , and all the samples X_i $(i = 1, 2, ...n)$ are drawn independently from this p-variate Gaussian distribution, then

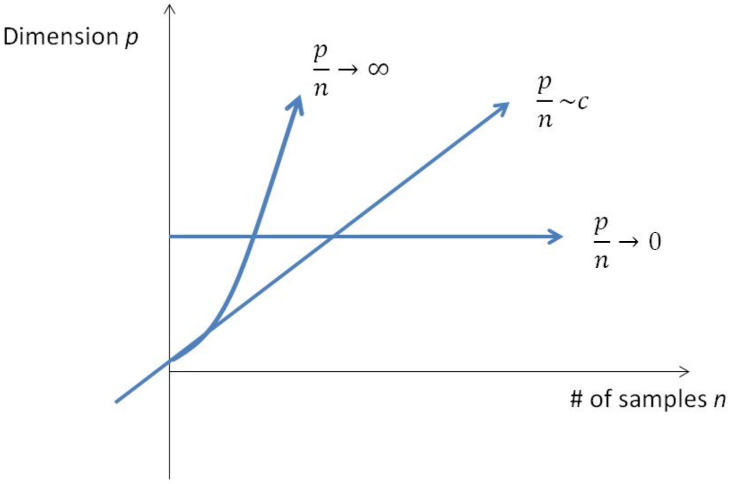

Figure 15.1. $\frac{p}{n}$ shows the relation between dimensionality and sample size

the matrix $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$ is also called a (random) *Wishart matrix*. And the *Wishart distribution* describes the probability distribution of this random Wishart matrix. In statistics, the *Wishart distribution* is a generalization to multiple dimensions of the chi-squared distribution, or, in the case of non-integer degrees of freedom, of the gamma distribution.

Theorem 15.1 (Marceenko-Pastur). *If n,$p \to \infty$, and $\frac{p}{n} \to c < 1$, then*

$$G_p(t) \quad = \quad \frac{1}{p}\#\{l_i : l_i \le nt\} \to G(t) \tag{15.14}$$

$$G'(t) \quad = \quad g(t) = \frac{\gamma}{2\pi t}\sqrt{(b-t)(t-a)}, \tag{15.15}$$

where $a = (1-\sqrt{c})^2, b = (1+\sqrt{c})^2$ and $G_p(t)$ describes the fraction of eigenvalues that is less than a particular nt.

In particular: all eigenvalues are contained in the interval: $[(1-\sqrt{c})^2, (1+\sqrt{c})^2]$.

One of the related results of the theorem is given by **Tracy-Widom Distribution**.

Theorem 15.2 (Tracy-Widom). *With appropriate centering and scaling, the logistic transform l_1 is approximately Tracy-Widom distributed.*

$$\frac{l_1 - \mu_{np}}{\sigma_{np}} \xrightarrow{d} W_1 \sim F_1 \tag{15.16}$$

where $\mu_{np} = (\sqrt{n-1} + \sqrt{p})^2$, $\sigma_{np} = (\sqrt{n-1} + \sqrt{p})(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{p}})^{\frac{1}{3}}$.

If we take the number of samples closes to the dimension (i.e, $n \approx p$), this says that standard deviations are of order $p^{\frac{1}{3}}$,instead of $p^{\frac{1}{2}}$.

More related results are given by *Wigner semi-circle law* and other random matrix ensembles.

The key idea here is that the top eigenvalue is l-Lipschitz and Lipschitz functions concentrate about their mean.

Back to the spiked covariance model: $y_i \sim \mathrm{N}(0, I + \sigma v v^T)$. If the number of samples $n \to \infty$ while p is fixed, then *Principle Component Analysis (PCA)*, a.k.a SVD of empirical covariance matrix $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T$ will recover $v, \forall \sigma > 0$.

What about for $\frac{p}{n} \to c < 1$? It turns out there is a phase transition.

Theorem 15.3. *For $\frac{p}{n} \to c < 1$, then $\hat{s_1}$ the largest eigenvalue of the empirical covariance matrix $\hat{\Sigma}$ behaves as follows:*

$$\hat{s_1} = \{ \begin{array}{ll} (1+\sqrt{c})^2 & \sigma_1 \leq 1+\sqrt{c} \\ \sigma_1 + \frac{c\sigma_1}{\sigma_1 - 1} & \sigma_1 > 1+\sqrt{c} \end{array} \tag{15.17}$$