Randomized Numerical Linear Algebra (RNLA) --- session I

Mohammadreza Soltani

These slides are based on “Lecture Notes on Randomized Linear Algebra” by Michael W. Mahoney
Introduction

• **Randomization and sampling** allow us to design provably accurate algorithms for problems that are
  - Massive
  - Computationally expensive or NP-hard
Introduction

• Basic algorithms for fundamental matrix problems
  - Matrix multiplication
  - Least-squares regression
  - Low-rank matrix approximation
Randomization meaning

• **Statistical approach**
  - observed data is a noisy/random version of ground truth

• **Algorithmic approach**
  - randomness is a computational resource for faster algorithms on a given observed data set
Why Randomization can be useful?

• Faster algorithms
• Simpler algorithms
• More interpretable algorithms and output
• Implicit regularization
First Problem

• Approximating the product of two matrices?
• If $A_{m \times n}$ and $B_{n \times p}$, we want to calculate $AB$

• *Traditional way*: 3 loop approach
• *RandNLA way*: Construct sketch from $A$ and $B$, and use them for multiplication

• $AB \approx CR$ where $C_{m \times c}$ and $R_{c \times p}$
The randomized sketches

• Random Sampling
  ➢ Multiplying the original matrix with a sparse matrix

• Random projections
  ➢ Multiplying the original matrix by a dense or nearly dense and consists of i.i.d. random variables
Randomized Algorithm for calculating $AB$

Algorithm The BasicMatrixMultiplication algorithm.

Input: An $m \times n$ matrix $A$, an $n \times p$ matrix $B$, a positive integer $c$, and probabilities $\{p_i\}_{i=1}^{n}$.

Output: Matrices $C$ and $R$ such that $CR \approx AB$

1: for $t = 1$ to $c$ do
2: Pick $i_t \in \{1, \ldots, n\}$ with probability $\Pr[i_t = k] = p_k$, in i.i.d. trials, with replacement
3: Set $C^{(t)} = A^{(i_t)}/\sqrt{cp_{i_t}}$ and $R^{(t)} = B^{(i_t)}/\sqrt{cp_{i_t}}$.
4: end for
5: Return $C$ and $R$. 
Lemma  Given matrices $A$ and $B$, construct matrices $C$ and $R$ with the BasicMatrixMultiplication algorithm. Then,

$$E[(CR)_{ij}] = (AB)_{ij}$$

and

$$\text{Var}[(CR)_{ij}] = \frac{1}{c} \sum_{k=1}^{n} \frac{A_{ik}^2 B_{kj}^2}{p_k} - \frac{1}{c} (AB)_{ij}^2.$$

Lemma  Given matrices $A$ and $B$, construct matrices $C$ and $R$ with the BasicMatrixMultiplication algorithm. Then,

$$E\left[\|AB - CR\|_F^2\right] = \sum_{k=1}^{n} \frac{\|A^{(k)}\|_2^2 \|B^{(k)}\|_2^2}{c p_k} - \frac{1}{c} \|AB\|_F^2.$$

Furthermore, if

$$p_k = \frac{\|A^{(k)}\|_2 \|B^{(k)}\|_2}{\sum_{k'=1}^{n} \|A^{(k')}\|_2 \|B^{(k')}\|_2},$$

then

$$E\left[\|AB - CR\|_F^2\right] = \frac{1}{c} \left(\sum_{k=1}^{n} \|A^{(k)}\|_2 \|B^{(k)}\|_2\right)^2 - \frac{1}{c} \|AB\|_F^2.$$
Lemma  Sampling probabilities \( \{p_i\}_{i=1}^{n} \) of the form
\[
p_k = \frac{\|A^{(k)}\|_2 \|B^{(k)}\|_2}{\sum_{k'=1}^{n} \|A^{(k')}\|_2 \|B^{(k')}\|_2},
\]
minimize \( \mathbb{E} \left[ \|AB - CR\|_F^2 \right] \).
Optimal and Nearly optimal sampling probability

\[ p_k = \frac{\|A^{(k)}\|_2 \|B^{(k)}\|_2}{\sum_{k'=1}^{n} \|A^{(k')}\|_2 \|B^{(k')}\|_2} \]

\[ p_k \geq \frac{\beta \|A^{(k)}\|_2 \|B^{(k)}\|_2}{\sum_{k'=1}^{n} \|A^{(k')}\|_2 \|B^{(k')}\|_2} \]

For \( 0 < \beta \leq 1 \)
Randomized Numerical Linear Algebra (RNLA) --- session 2

Mohammadreza Soltani
Sample complexity

First and simple way to obtain how many column-row pair ($c$) we need:

**Lemma (Markov’s Inequality)** Let $X$ be a real-valued random variable such that $X \geq 0$. Then, $\forall a \geq 0$, $\Pr [X \geq a] \leq \frac{E[X]}{a}$.

**Theorem**: If we select $c \geq \frac{1}{\beta \delta^2 \epsilon^2}$ then with probability $1 - \delta$, we have:

$$||AB - CR||_F \leq \epsilon ||A||_F ||B||_F$$
Concentration inequality for Random Variables

Lemma  (Vanilla Chernoff, other direction) \( \forall t < 0: \Pr [X \leq a] = \Pr [e^{tX} \leq e^{ta}] \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}}. \)

In particular, \( \Pr [X \leq a] \leq \min_{t>0} \frac{\mathbb{E}[e^{tX}]}{e^{ta}}. \)

There are a lot of variants of this basic result, depending on what is known about the given distribution, how tight a bound one can provide for \( \mathbb{E}[e^{tX}] \), etc. Here are two versions.

Theorem  (Hoeffding) Let \( \{X_i\}_{i=1}^n \) be r.v. such that \( X_i \in [a_i, b_i] \), for all \( i \), and let \( X = \sum_{i=1}^n X_i. \) Then

\[
\Pr [|X - \mathbb{E}[x]| \geq t] \leq 2 \exp \left( \frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right)
\]

Theorem  (Bernstein) Let \( \{X_i\}_{i=1}^n \) be r.v. such that \( \mathbb{E}[X], \mathbb{E}[X^2] = \sigma^2, |X| \leq M, X_i \) are i.i.d. copies. Then, for all \( t > 0, \)

\[
\Pr [|X - \mathbb{E}[x]| \geq t] \leq \exp \left( \frac{-t^2}{2n\sigma^2 + \frac{4}{3}tM} \right)
\]
Dependency between random Variables

Definition  A sequence of random variables $Z_0, Z_1, \ldots$ is a martingale with respect to a sequence $X_0, X_1, \ldots$ if, \( \forall n \geq 0, \)

- \( Z_n = Z_n(X_0, X_1, \ldots, X_n) \), i.e., it is a function of the \( X_i \)
- \( \mathbb{E}[|Z_n|] < \infty \)
- \( \mathbb{E}[Z_{n+1}|X_0 \cdots X_n] = Z_n \)

A sequence is a martingale if it is a martingale with respect to itself.

Theorem  (Azuma-Hoeffding)  Let $X_0, \ldots, X_n$ be a martingale such that $|X_k - X_{k-1}| \leq c_k$. Then, \( \forall t \geq 0, \lambda > 0, \)

\[
\Pr[|X_t - X_0| \geq \lambda] \leq 2 \exp \left( \frac{-\lambda^2}{\sum_{k=1}^{t} c_k^2} \right).
\]

As a corollary, if $|X_k - X_{k-1}| \leq c$, then

\[
\Pr \left[ |X_t - X_0| \geq \lambda c \sqrt{t} \right] \leq 2 \exp \left( -\frac{\lambda^2}{2} \right).
\]
Theorem  Suppose $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, $c \in \mathbb{Z}^+$ such that $1 \leq c \leq n$, and $\{p_i\}_{i=1}^n$ are such that $\sum_{i=1}^n p_i = 1$ and such that for some positive constant $\beta \leq 1$

$$p_k \geq \frac{\beta \left\| A^{(k)} \right\|_2 \left\| B^{(k)} \right\|_2}{\sum_{k'=1}^n \left\| A^{(k')} \right\|_2 \left\| B^{(k')} \right\|_2}.$$ 

Construct $C$ and $R$ with the BasicMatrixMultiplication algorithm, and let $CR$ be an approximation to $AB$. Then,

$$\mathbb{E} \left[ \left\| AB - CR \right\|_F^2 \right] \leq \frac{1}{\beta c} \left\| A \right\|_F^2 \left\| B \right\|_F^2.$$ 

Furthermore, let $\delta \in (0, 1)$ and $\eta = 1 + \sqrt{\frac{8}{\beta}} \log(1/\delta)$. Then, with probability at least $1 - \delta$,

$$\left\| AB - CR \right\|_F^2 \leq \frac{\eta^2}{\beta c} \left\| A \right\|_F^2 \left\| B \right\|_F^2.$$
If \( B = A^T \)

**Theorem** Suppose \( A \in \mathbb{R}^{m \times n}, c \in \mathbb{Z}^+, 1 \leq c \leq n, \) and \( \{p_i\}_{i=1}^n \) are such that \( \sum_{i=1}^n p_i = 1 \) and such that \( p_k \geq \frac{\beta\|A^{(k)}\|^2_2}{\|A\|^2_F} \) for some positive constant \( \beta \leq 1 \). Furthermore, let \( \delta \in (0,1) \) and \( \eta = 1 + \sqrt{(8/\beta) \log(1/\delta)} \). Construct \( C \) (and \( R = C^T \)) with the BasicMatrixMultiplication algorithm, and let \( CC^T \) be an approximation to \( AA^T \). Then,

\[
\mathbb{E}[\|AA^T - CC^T\|_F] \leq \frac{1}{\sqrt{\beta c}} \|A\|_F^2
\]

and with probability at least \( 1 - \delta \),

\[
\|AA^T - CC^T\|_F \leq \frac{\eta}{\sqrt{\beta c}} \|A\|_F^2.
\]
What if we want to drive approximation error in Spectral norm

• We need to apply matrix version of concentration inequality

• A brief history:
  
  • Alswede-Winter: the original result related to bounding the matrix m.g.f. that started this recent flurry of work in this area.
  
  • Christofides-Markstron: introduced a matrix version of Hoeffding-Azuma
  
  • Rudelson and Vershynin: had the original bounds for operator-valued random variables that were originally used to bound the spectral norm error. They bounds had a similar form, but they depended on heavier-duty arguments from convex analysis, and they sometimes didn’t provide constants, which made it awkward for numerical implementation.
  
  • Gross, Recht, and Oliviera: several different versions of matrix Chernoff bounds.
  
  • Tropp: provides a nice review of this line of work.
Two versions

• Oliviera

Lemma. \( \text{Let } X_1, \ldots, X_n \text{ be i.i.d. random column vectors in } \mathbb{R}^d \text{ such that } \|X_i\|_2 \leq M \text{ a.s. and } \|E[X_i X_i^*]\|_2 \leq 1. \text{ Then, } \forall t \geq 0, \)

\[
\Pr \left[ \left\| \frac{1}{n} \sum_{i=1}^{n} X_i X_i^* - E[X_1 X_1^*] \right\|_2 \geq t \right] \leq (2n)^2 \exp \left( \frac{-nt^2}{16M^2 + 8M^2t} \right).
\]

• Recht

Theorem. (Noncommutative Bernstein Inequality) \( \text{Let } X_1, \ldots, X_L \text{ be independent zero-mean random matrices of dimension } d_1 \times d_2. \text{ Suppose that } \rho_k = \max\{ \|E[X_k X_k^*]\|_2, \|E[X_k^* X_k]\|_2 \} \text{ and that } \|X_k\|_2 \leq M \text{ a.s., for all } k. \text{ Then, } \forall \tau > 0, \)

\[
\Pr \left[ \left\| \sum_{k=1}^{L} X_k \right\|_2 > \tau \right] \leq (d_1 + d_2) \exp \left( \frac{-\tau^2/2}{\sum_{k=1}^{L} \rho_k^2 + M\tau/2} \right).
\]
Approximation error

Theorem  Let $A \in \mathbb{R}^{m \times n}$, and consider approximating $AA^T$. Construct a matrix $C \in \mathbb{R}^{m \times c}$, consisting of $c$ sampled and rescaled columns of $A$, with the BasicMatrixMultiplication algorithm, where the sampling probabilities $\{p_i\}_{i=1}^n$ satisfy

$$p_i \geq \beta \frac{\|A^{(i)}\|_2^2}{\|A\|_F^2}$$

for all $i \in [n]$, and for some constant $\beta \in (0,1]$. Assume, for simplicity, that $\|A\|_2 \leq 1$ and $\|A\|_F^2 \geq 1/24$, let $\epsilon \in (0,1)$ be an accuracy parameter, and let

$$c \geq \frac{96 \|A\|_F^2}{\beta \epsilon^2} \log \left( \frac{96 \|A\|_F^2}{\beta \epsilon^2 \sqrt{\delta}} \right).$$

Then, with probability $\geq 1 - \delta$, we have that

$$\|AA^T - CC^T\|_2 \leq \epsilon.$$
Sketching through random projection

• Relating to Johnson-Lindenstrauss transform

we have $n$ points $\{u_i\}_{i=1}^n$, each of which is in $\mathbb{R}^d$, e.g., the rows of an $n \times d$ matrix $A$, and we want to find $n$ points $\{v_i\}_{i=1}^n$, each of which is in $\mathbb{R}^k$ such that

• $k \ll d$

• $||v_i|| \approx ||u_i||$, for all $i$

• $||v_i - v_{i'}|| \approx ||u_i - u_{i'}||$, for all $i, i'$

Lemma (JL lemma) Given $n$ points $\{u_i\}_{i=1}^n$, each of which is in $\mathbb{R}^d$, $P \in \mathbb{R}^{d \times k}$ be such that $P_{ij} = \frac{1}{\sqrt{k}}N(0, 1)$, and let $\{v_i\}_{i=1}^n$ be points in $\mathbb{R}^k$ defined as $v_i = u_i P$. Then, if $k \geq \frac{9 \log(n)}{\epsilon^2 - \epsilon^3}$, for some $\epsilon \in (0, 1/2)$, then with probability at least $1/2$, all pairwise distances are preserved, i.e., for all $i, i'$, we have

$$(1 - \epsilon) \frac{||u_i - u_{i'}||_2^2}{2} \leq \frac{||v_i - v_{i'}||_2^2}{2} \leq (1 + \epsilon) \frac{||u_i - u_{i'}||_2^2}{2}.$$
Two families of random projection for random projection

\[ P_{ij} = \begin{cases} 
1/\sqrt{k} & \text{w.p. } = 1/2 \\
-1/\sqrt{k} & \text{w.p. } = 1/2 
\end{cases} \]

\[ P_{ij} = \begin{cases} 
3/\sqrt{k} & \text{w.p. } = 1/6 \\
0 & \text{w.p. } = 2/3 \\
-3/\sqrt{k} & \text{w.p. } = 1/6 
\end{cases} \]
A random projection algorithm for approximating matrix multiplication

Given as input an $m \times n$ matrix $A$, an $n \times p$ matrix $B$, a positive integer $c$, do the following.

1. Let $\Pi$ be an $n \times c$ random projection matrix, as defined above.

2. Let $C = A\Pi$ and $R = \Pi^T B$ be sketches of the columns of $A$ and rows of $B$.

3. Compute and return $CR = A\Pi\Pi^T B$. 
Summary

• The matrix multiplication primitive can be done in one to two ways:
  - In a data-aware manner, in which we perform random sampling with sampling probabilities that depend on the input matrices
  - In a data-agnostic manner, in which we perform random projections without looking at the input data.