EE270
Large scale matrix computation, optimization and learning

Instructor : Mert Pilanci

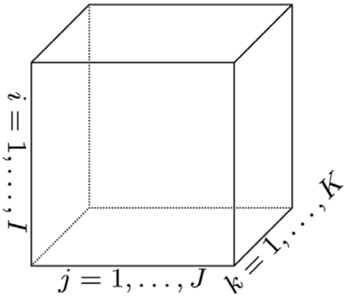
Stanford University

Thursday, Jan 16 2020

Randomized Linear Algebra
Lecture 4: Approximate Tensor Products, Randomized Verification and Concentration Inequalities

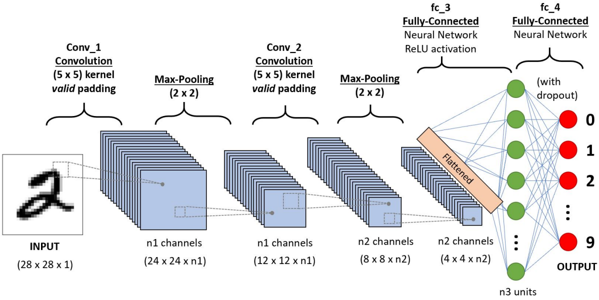
Tensors and tensor multiplication

- A tensor is a multidimensional array
- Order of a tensor: number of dimensions, also known as modes
- An element (i, j, k) of a third-order tensor X is denoted by $X_{i,j,k}$
- (Frobenious) norm of a tensor $\|X\|_F = \sqrt{\sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} |X_{i_1 i_2 \cdots i_N}|^2}$



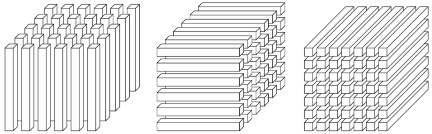
Tensors and tensor multiplication

- Deep Neural Network weights and activations are typically tensors

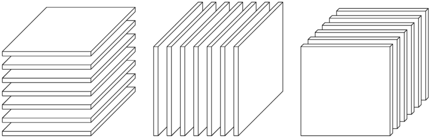


Tensors and tensor multiplication

- Fibers are the higher-order analogue of matrix rows and columns. Defined by fixing every index but one
- Slices are two-dimensional sections of a tensor, defined by fixing all but two indices



(a) Mode-1 (column) fibers: $\mathbf{x}_{:jk}$ (b) Mode-2 (row) fibers: $\mathbf{x}_{i:k}$ (c) Mode-3 (tube) fibers: $\mathbf{x}_{ij:}$



(a) Horizontal slices: $\mathbf{X}_{i::}$ (b) Lateral slices: $\mathbf{X}_{:j:}$ (c) Frontal slices: $\mathbf{X}_{::k}$ (or $\mathbf{X}_k$)

Tensor n-Mode Product

- n-mode (matrix) product of a tensor $A \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ with a matrix $B \in \mathbb{R}^{p \times d_n}$ is elementwise

$$(A \quad \times_n \quad B)_{i_1, \cdots, i_{n-1}\, j\, i_{n+1} \cdots i_N} = \sum_{i_n=1}^{d_n} A_{i_1 i_2 \cdots i_n \cdots d_N} B_{j i_n}$$

- each mode-n fiber of A is multiplied by the matrix B

Approximate Tensor Multiplication

Algorithm 1 Approximate Tensor n-Mode Product via Sampling

Input: An $d_1 \times \cdots \times d_n \times \cdots \times d_N$ dimensional tensor A and an $p \times d_n$ dimensional tensor B, an integer m and probabilities $\{p_k\}_{k=1}^{d_n}$

Output: Tensors CR such that $CR \approx AB$

for $t = 1$ to m **do**

  Pick $i_t \in \{1, \ldots, d_n\}$ with probability $\mathbb{P}[i_t = k] = p_k$ in i.i.d. with replacement

  Set $C^{(t)} = \frac{1}{\sqrt{mp_{i_t}}} A_{:,i_t,:}$ and $R_{(t)} = \frac{1}{\sqrt{mp_{i_t}}} B_{:,i_t,:}$

end for

- We can multiply CR using the classical algorithm
- Complexity $O(d_1 \cdots d_{n-1} m d_n \cdots d_N p)$

Approximate Tensor Multiplication: Mean and variance

$$M_{\vec{i}j} \triangleq (A \quad \times_n \quad B)_{i_1,\cdots,i_{n-1}\,j\,i_{n+1}\cdots i_N} = \sum_{i_n=1}^{d_n} A_{i_1 i_2\cdots i_n\cdots i_N} B_{j i_n}$$

$$\hat{M}_{\vec{i}j} \triangleq \sum_{i_n=1}^{m} \frac{1}{p_{i_n}} A_{i_1 i_2\cdots i_n\cdots i_N} B_{j i_n}$$

- Mean and variance of the matrix multiplication estimator

  Lemma
- $\mathbb{E}\left[\hat{M}_{\vec{i}j}\right] = M_{\vec{i}j}$
- **Var** $\left[\hat{M}_{\vec{i}j}\right] = \frac{1}{m}\sum_{i_n=1}^{d_n} \frac{1}{p_{i_n}} A^2_{i_1 i_2\cdots i_n\cdots i_N} B^2_{j i_n} - \frac{1}{m}(M_{\vec{i}j})^2$

Approximate Tensor Multiplication: Mean and variance

$$M_{\vec{i}j} \triangleq (A \quad \times_n \quad B)_{i_1,\cdots,i_{n-1}\, j\, i_{n+1}\cdots i_N} = \sum_{i_n=1}^{d_n} A_{i_1 i_2\cdots i_n\cdots i_N} B_{j i_n}$$

$$\hat{M}_{\vec{i}j} \triangleq \sum_{i_n=1}^{m} \frac{1}{p_{i_n}} A_{i_1 i_2\cdots i_n\cdots i_N} B_{j i_n}$$

- Mean and variance of the matrix multiplication estimator

  Lemma
- $\mathbb{E}\left[\hat{M}_{\vec{i}j}\right] = M_{\vec{i}j}$
- **Var** $\left[\hat{M}_{\vec{i}j}\right] = \frac{1}{m}\sum_{i_n=1}^{d_n} \frac{1}{p_{i_n}} A^2_{i_1 i_2\cdots i_n\cdots i_N} B^2_{j i_n} - \frac{1}{m}(M_{\vec{i}j})^2$
- $\text{minimize}_p\, \mathbb{E}\|\hat{M} - M\|_F^2 = \sum_{\vec{i}j}$ **Var** $\left[\hat{M}_{\vec{i}j}\right]$

Approximate Multiplication for Tensors

$$\hat{M}_{\vec{i}j} \triangleq \sum_{i_n=1}^{m} \frac{1}{p_{i_n}} A_{i_1 i_2 \cdots i_n \cdots i_N} B_{j i_n}$$

- Importance sampling distribution

$$p_k = \frac{\|A_{:\cdots k \cdots :}\|_F \|B_{:k}\|_F}{\sum_k \|A_{:\cdots k \cdots :}\|_F \|B_{:k}\|_F}$$

Verifying Matrix Multiplication

- Given three $n \times n$ matrices A, B, M
- verify whether
$$AB = M$$
- Naive method: $O(n^3)$

Randomized Algorithm for Verifying Matrix Multiplication

- Sample a random vector $r = [r_1, ..., r_n]^T$
- Compute ABr by first computing Br and then $A(Br)$
- Compute Mr
- If $A(Br) \neq Mr$, then $AB \neq M$
- Otherwise, return $AB = M$

Randomized Algorithm for Verifying Matrix Multiplication

- Sample a random vector $r = [r_1, ..., r_n]^T$
- Compute ABr by first computing Br and then $A(Br)$
- Compute Mr
- If $A(Br) \neq Mr$, then $AB \neq M$
- Otherwise, return $AB = M$
- Complexity: three matrix-vector multiplications $O(n^2)$ Freivalds' Algorithm (1977)

Failure Probability

- Let $r = [r_1, ..., r_n]^T$ be i.i.d. $+1, -1$ each with probability $\frac{1}{2}$
- Lemma $\mathbb{P}[ABr = Mr] \leq \frac{1}{2}$

Multiple trials

- $r = [r_1, ..., r_n]^T$ be i.i.d. $0, 1$ each with probability $\frac{1}{2}$ also works
- To improve the error probability, we run the algorithm independently k times with

  $r_1, ..., r_k \in \mathbb{R}^n$ i.i.d.
- If we ever find an r_k such that

  $ABr_k \neq Mr$
- then the algorithm correctly returns $AB \neq M$

Multiple trials

- $r = [r_1, ..., r_n]^T$ be i.i.d. $0, 1$ each with probability $\frac{1}{2}$ also works
- To improve the error probability, we run the algorithm independently k times with

  $r_1, ..., r_k \in \mathbb{R}^n$ i.i.d.
- If we ever find an r_k such that

  $ABr_k \neq Mr$
- then the algorithm correctly returns $AB \neq M$
- If we always find $ABr = Mr$, then the error probability is at most $\frac{1}{2^k}$
- For $k = 25$ we have error probability $\leq 10^{-9}$.

Concentration bounds: Tighter success probability

- In AMM size of the sample is $m = \frac{1}{\delta \epsilon^2}$.
  dependence on the failure probability δ is not ideal
  we can do better
- recall Markov's Inequality
  For $Z > 0$ and $t > 0$

$$\mathbb{P}[Z > a] \leq \frac{\mathbb{E}Z}{a}$$

- Chebyshev's inequality
  Let X be a random variable with expectation $\mathbb{E}[X]$ and variance $\mathbf{Var}[X]$

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq \frac{\mathbf{Var(X)}}{t^2}.$$

Concentration of independent sums

- Chernoff Bound[1]
- Let $X_1, ..., X_m$ be independent random variables $\in [0,1]$ and let $\mu = \mathbb{E}X_1$

$$\mathbb{P}[\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| > t\mu] \leq 2e^{-m\frac{t^2\mu}{3}}$$

[1] There are other versions of the Chernoff bound which have better constants

Application 1: Monte Carlo Approximations

- Estimating π
- Sample $z_1, ..., z_m$ i.i.d. uniform in $[0, 1]^2$
- Let $Z_i = 1$ if $\|z_i\|_2 \leq 1$ and 0 otherwise
- $\mathbb{P}[Z_i = 1] = \frac{\pi}{4}$

Application 1: Monte Carlo Approximations

- Estimating π
- Sample $z_1, ..., z_m$ i.i.d. uniform in $[0,1]^2$
- Let $Z_i = 1$ if $\|z_i\|_2 \leq 1$ and 0 otherwise
- $\mathbb{P}[Z_i = 1] = \frac{\pi}{4}$
- Applying Chernoff bound we get

$$\left| \frac{1}{m} \sum_{i=1}^{m} Z_i - \frac{\pi}{4} \right| \leq \epsilon \frac{\pi}{4}$$

  with probability at least $1 - 2e^{-m\epsilon^2 \frac{\pi}{12}}$
- we can pick $m \geq \frac{12}{\pi\epsilon^2} \log \frac{2}{\delta}$ and obtain an estimate $\hat{\pi}$ such that $(1-\epsilon)\pi \leq \hat{\pi} \leq (1+\epsilon)\pi$ with probability at least $1 - \delta$
  the range $[(1-\epsilon)\pi, (1+\epsilon)\pi]$ is a confidence interval

Application 2: Amplifying Probability of Success

- Suppose we have a randomized algorithm which produces an ϵ approximation $|\hat{x} - x^*| \leq \epsilon$

  with probability at least 0.9
- Repeat the algorithm m times independently
- Take median of m outputs

Application 2: Amplifying Probability of Success

- Suppose we have a randomized algorithm which produces an ϵ approximation $|\hat{x} - x^*| \leq \epsilon$

  with probability at least 0.9
- Repeat the algorithm m times independently
- Take median of m outputs
- Let $X_i = 1$ if the i-th trial is **good**, i.e., $|\hat{x}_i - x^*| \leq \epsilon$
- Median of the m outputs is also **good**, i.e., $|\text{Median}(\hat{x}_i) - x^*| \leq \epsilon$ if **at least half** of the X_i's are one
- Chernoff Bound implies that $\left|\frac{1}{m}\sum_{i=1}^{m} X_i - 0.9\right| \leq 0.9t$ with probability $1 - e^{-t^2 0.9m/3}$. Pick $t = 0.4/0.9$
- Median is an ϵ approximation with probability at least $1 - e^{-0.059m}$

  e.g., for $m = 200$, failure probability is $\leq 7 \times 10^{-6}$.

"Median" for Approximate Matrix Multiplication

- Chernoff bound implies that majority of estimators are good
- The definition of median does not extend to the matrix case in a simple way
- Recall AMM final probability bound

  For any $\delta > 0$, set $m = \frac{1}{\delta\,\epsilon^2}$ to obtain

$$\mathbb{P}\left[\|AB - CR\|_F > \epsilon \|A\|_F \|B\|_F\right] \leq \delta$$

- suppose $\|A\|_F = \|B\|_F = 1$ and let $\epsilon = 0.1$, $\delta = 0.9$
- Repeat independently and obtain $C_1 R_1, ..., C_t R_t$ in t independent trials

  $\|AB - C_i R_i\|_F < 0.1$ with probability 0.9 for each i

"Median" for Approximate Matrix Multiplication

- Repeat independently and obtain $C_1R_1, ..., C_tR_t$ in t independent trials

  $\|AB - C_iR_i\|_F < 0.1$ with probability 0.9 for each i
- we don't know which ones are **good**, i.e., $\|AB - C_iR_i\|_F < 0.1$

"Median" for Approximate Matrix Multiplication

- Repeat independently and obtain $C_1R_1, ..., C_tR_t$ in t independent trials

  $\|AB - C_iR_i\|_F < 0.1$ with probability 0.9 for each i
- we don't know which ones are **good**, i.e., $\|AB - C_iR_i\|_F < 0.1$
- Let $X_i = 1$ if the i-th trial is **good** and $X_i = 0$ otherwise
- Chernoff Bound implies that $\frac{1}{m}\sum_{i=1}^{m} X_i \geq 0.5$ with probability $1 - e^{-0.059m}$, i.e., **at least half of the matrices are good**

"Median" for Approximate Matrix Multiplication

- Repeat independently and obtain $C_1R_1, \ldots, C_tR_t$ in t independent trials

  $\|AB - C_iR_i\|_F < 0.1$ with probability 0.9 for each i
- we don't know which ones are **good**, i.e., $\|AB - C_iR_i\|_F < 0.1$
- Let $X_i = 1$ if the i-th trial is **good** and $X_i = 0$ otherwise
- Chernoff Bound implies that $\frac{1}{m}\sum_{i=1}^{m} X_i \geq 0.5$ with probability $1 - e^{-0.059m}$, i.e., **at least half of the matrices are good**
- Compute $\rho_i \triangleq |\{j \mid j \neq i,\ \|C_iR_i - C_jR_j\|_F \leq 0.2\}|$
- Output C_kR_k such that $\rho_k \leq \frac{t}{2}$
- **Lemma:** $\|AB - C_kR_k\|_F \leq 0.3$ with probability at least$1 - e^{-0.059m}$.

Median Trick for Matrices

- **Proof:**
- triangle inequality: $\|X+Y\|_F \leq \|X\|_F + \|Y\|_F$ and
- reverse triangle inequality: $\|X+Y\|_F \geq \|X\|_F - \|Y\|_F$
- for matrices $X, Y \in \mathbb{R}^{n \times p}$ imply

  $\|C_i R_i - C_j R_j\|_F \leq \|C_i R_i - AB\|_F + \|C_j R_j - AB\|_F$

  $\|C_i R_i - C_j R_j\|_F \geq \|C_i R_i - AB\|_F - \|C_j R_j - AB\|_F$

Median Trick for Matrices

- **Proof:**
- triangle inequality: $\|X+Y\|_F \leq \|X\|_F + \|Y\|_F$ and
- reverse triangle inequality: $\|X+Y\|_F \geq \|X\|_F - \|Y\|_F$
- for matrices $X, Y \in \mathbb{R}^{n \times p}$ imply
  $\|C_i R_i - C_j R_j\|_F \leq \|C_i R_i - AB\|_F + \|C_j R_j - AB\|_F$
  $\|C_i R_i - C_j R_j\|_F \geq \|C_i R_i - AB\|_F - \|C_j R_j - AB\|_F$
- If $C_i R_i$ is **good**,$\|AB - C_i R_i\|_F \leq 0.1$ then
  it is close to at least half of the other $C_j R_j$'s
  $\rho_i \triangleq |\{j \mid j \neq i, \ \|C_i R_i - C_j R_j\|_F \leq 0.2\}| \geq \frac{t}{2}$ by triangle inequality

Median Trick for Matrices

- **Proof:**
- triangle inequality: $\|X + Y\|_F \leq \|X\|_F + \|Y\|_F$ and
- reverse triangle inequality: $\|X + Y\|_F \geq \|X\|_F - \|Y\|_F$
- for matrices $X, Y \in \mathbb{R}^{n \times p}$ imply
  $\|C_i R_i - C_j R_j\|_F \leq \|C_i R_i - AB\|_F + \|C_j R_j - AB\|_F$
  $\|C_i R_i - C_j R_j\|_F \geq \|C_i R_i - AB\|_F - \|C_j R_j - AB\|_F$
- If $C_i R_i$ is **good**, $\|AB - C_i R_i\|_F \leq 0.1$ then
  it is close to at least half of the other $C_j R_j$'s
  $\rho_i \triangleq |\{j \mid j \neq i, \ \|C_i R_i - C_j R_j\|_F \leq 0.2\}| \geq \frac{t}{2}$ by triangle inequality
- If $C_i R_i$ is **bad**, i.e., $\|AB - C_i R_i\|_F > 0.3$ then
  $\|C_i R_i - C_j R_j\|_F \geq 0.2$ by triangle inequality and $\rho_i \leq \frac{t}{2}$

Questions?