Randomized Algorithms: Finding the product of two $n \times n$ matrices in $O(n^2)$ time

Amit Agarwal's scribe for Prof. Ravi Kannan's first guest lecture

January 21, 2007

1 Introduction

This is the first in the series of lectures on randomized algorithms that have as input, large data sets. We consider the problem of multiplying two $n \times n$ matrices. The algorithm given my Solovay and Strassen solves the problem exactly in $O(n^{log_2 7})$ time and there have been many further improvements. But our aim here is to develop faster (though error-prone) algorithms for this problem. Sampling is used to obtain some estimate of the values. As a warm up, consider the problem of finding the sum of n real numbers by looking at only a few of them. More specifically, we perform c independent and identical trials. In the t^{th} trial we pick a number uniformly from the set $1 \dots n$. (One issue here is whether the trials are done with replacement or without replacement. Trials with replacement offer the advantage of simple mathematics, whereas one may achieve better results in trials without replacement especially if the number c of trials is not too small compared to n. In fact, the following principle can be proved under mild conditions:

Fact 1. *Sampling without replacement is no worse than sampling with replacement.*

Here we are picking elements with replacement. [Picking elements with replacement or without replacement does not make any difference if the number of elements picked is small]. Let us use a_{i_t} to denote the value of the element picked up in the t^{th} trial. Let us define random variables X_t and X as follows

$$\begin{aligned}
X_t &= \frac{n}{c} a_{i_t} \\
X &= \sum_{t=1}^{c} \frac{n}{c} a_{i_t} \\
E[X_t] &= \frac{n}{c} \sum_{i=1}^{n} \frac{1}{n} a_i \\
&= \frac{1}{c} a \\
E[X] &= a
\end{aligned}$$

Thus X is an "unbiased estimate" of a, meaning simply that the expected value of X is equal to a.

Fact 2. *If Y is the sum of independent random variables $Y_1, \ldots, Y_t$, then Var[Y] $= \sum_{i=1}^{t} Var[Y_i]$.*

Proof. Left as an exercise to the reader. □

Now,

$$\begin{aligned}
Var[X] &= \sum Var[X_t] \\
&= cVar[X_1]
\end{aligned}$$

$$\begin{aligned}
Var[X] &= c \sum_{i=1}^{n} \frac{1}{n} \left(\frac{n}{c} a_i - \frac{a}{c}\right)^2 \\
&= \frac{1}{c} \left(n \sum_{i=1} n a_i^2 - a^2 \right)
\end{aligned}$$

Aside: For an intuitive understanding of p_i, consider the case when all $|a_i|$ are $\Theta(M)$ for some real M. Then $E[X] = \Theta(NM)$. $Var[x] \leq \frac{1}{c} \sum_{i=1} n n a_i^2$ is $\frac{1}{c} O(n^2 m^2)$ and the standard deviation is $\sqrt{Var} = O(\frac{1}{\sqrt{c}} nm)$. If we want to achieve relative error $\varepsilon \geq 0$ in estimating $E[X]$, we want the standard deviation to be $\leq \varepsilon E[X]$. This is achieved by taking $c \in \Omega(\frac{1}{c^2})$.

Remark: We may also consider a situation where each element decides on its own to be part of the set of c elements or not. Suppose each element does this with probability $\frac{c}{n}$. Then this is roughly equivalent to the method discussed above, provided $c \leq\leq n$.

So far we had picked elements with equal probability. Now let us see what vector of probabilities of picking elements would give us lower error. For this purpose let us define p_i to be the probability of picking the i^{th} element such that, $\sum_{i=1}^{n} p_i = 1$. Let us define,

$$\begin{aligned}
Y_t &= \frac{1}{c}\frac{a_{i_t}}{p_{i_t}} \\
Y &= \sum_{t=1}^{c} Y_t \\
E[Y] &= cE[Y_1] \\
&= \sum_{i=1}^{n} \frac{a_i p_i}{p_i} \\
&= a
\end{aligned}$$

(Note that here $a = \sum_{i=1} na_i$.)

$$Var[Y] = cVar[Y_1]$$

because all the trials are independent and identical.

$$\begin{aligned}
Var[Y] &= \frac{1}{c}\sum_{i=1}^{n} \frac{a_i^2}{p_i^2} p_i - a^2 \\
&= \frac{1}{c}\sum_{i=1}^{n} \frac{a_i^2}{p_i} - a^2 \\
&= \frac{1}{c}\sum_{i=1}^{n} \frac{a_i^2}{p_i}
\end{aligned}$$

(Exercise) Prove that the minimum of the above variance is achieved when $p_i = \frac{|a_i|}{\sum_{j=1}^{n} |a_j|}$ (assuming that all the a_i's are non-zero.

Proof. (Hint) Use calculus. □

For p_i as given above, $Var[Y] \le \frac{1}{c}(\sum_{j=1}^{n} |a_j|)^2$.
On the other hand, for the uniform case,
$Var[X] \le \frac{1}{c}\sum_{j=1}^{n} a_j^2$.

Theorem 1. *(Cauchy-Schawz's inequality) For any reals $a_1,\dots,a_n$ and $b_1,\dots,b_n$,*
$\|\sum_{i=1}^{n} a_i b_i\| \le (\sum_{i=1}^{n} a_i^2)^{\frac{1}{2}} (\sum_{i=1}^{n} b_i^2)^{\frac{1}{2}}$.

By the Cauchy-Schawz inequality,

$$(\sum_{i=1} n\|a_j\|)^2 \le n \sum_{i=1} na_j^2$$

Hence, the variance of the weighted sampling method is no worse than the uniform method.

But the problem in using these p_i's is that finding them would be as much work as in finding $\sum_{j=1}^{n} a_j$. However, this will be useful for matrix multiplication.

In the case of matrices, each a_i corresponds to a column(or row) and we will be able to compute all the p_i's in $O(n^2)$ time.

2 Matrix Multiplication

Now let us turn our attention to the problem we originally set out to attack, namely approximate randomized algorithms for matrix multiplication. We consider two $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$. The $(i,j)^{th}$ element of the product $(\mathbf{AB})_{ij}$ is given by $\sum_{k=1}^{n}(\mathbf{A}_{ik})(\mathbf{B}_{kj})$.

A first thought: Consider each element of the product as the sum of n real numbers and apply the weighted random technique developed above. As a first idea we could apply random sampling to find the sum for each i, j separately. However, in applying this sampling, we would need $O(n^3)$ time. Since, for each (i,j) we seem to need to read up $O(n)$ entries, etc.

Notation: $\mathbf{A}^{(i)} = i^{th}$ column vector of $\mathbf{A}$. $\mathbf{B}_{(i)} = i^{th}$ row vector of $\mathbf{b}$. Let $p_i : \sum_{i=1}^{n} p_i = 1$ denote the probability of choosing the i^{th} column (or row). Assume for now that these are given.

In the t^{th} trial pick $i_t \in 1,\ldots,n$ with Prob $(i_t = i) = p_i$. Define matrices $\mathbf{C}$ and $\mathbf{R}$: $\frac{\mathbf{A}^{(i_t)}}{\sqrt{cp_{i_t}}}$ is the t^{th} column of $\mathbf{C}$. $\frac{\mathbf{B}_{(i_t)}}{\sqrt{cp_{i_t}}}$ is the t^{th} row of $\mathbf{R}$. Then we will show that $\mathbf{CR}$ is the approximation of $\mathbf{AB}$.

We now state lemmas dealing with the errors in the above approximation process.

Lemma 2. $E(\mathbf{CR}) = \mathbf{AB}$ *(shorthand for* $E(\mathbf{CR}_{ij}) = (\mathbf{AB})_{ij} \forall i, j.$*)*

Proof. $\mathbf{CR}_{ij} = \sum_{t=1}^{c} \mathbf{C}_{it}\mathbf{R}_{tj}$. Denote $\mathbf{C}_{it}\mathbf{R}_{tj}$ as X_t. The X_t's are independent and identically distributed, so that

$$\begin{aligned} E(\sum_{t=1}^{c} X_t) &= c\sum_{t=1}^{c} X_1 = cE(\mathbf{C}_{ii_1}\mathbf{R}_{i_1 j}) \\ &= c\frac{E(\mathbf{A}_{ii_1}\mathbf{B}_{i_1 j})}{cp_{i_1}} \end{aligned}$$

$i_1 = k$ with probability p_k.

$$\begin{aligned} E(\sum_{t=1}^{c} X_t) &= c\sum_{k=1}^{n} \frac{p_k \mathbf{A}_{ik}\mathbf{B}_{kj}}{cp_k} \\ &= (\mathbf{AB})_{ij} \end{aligned}$$

□

Lemma 3. $Var[\mathbf{CR}_{ij}] = \frac{1}{c}\sum_{k=1} n\frac{\mathbf{A}_{ik}^2\mathbf{B}_{kj}^2}{p_k} - \frac{1}{c}(\mathbf{AB})_{ij}^2$.

Proof.

$$\begin{aligned} Var[\sum_{t=1}^{c} X_t] &= cVar[X_1] \\ &= cE((\mathbf{C}_{i1}\mathbf{R}_{1j})^2) - c(E(\mathbf{C}_{i1}\mathbf{R}_{1j}))^2 \\ &= c\sum_{k=1} n\frac{p_k \mathbf{A}_{ik}^2\mathbf{B}_{kj}^2}{c^2 p_k^2} - \frac{1}{c}(\mathbf{AB})_{ij}^2 \end{aligned}$$

□

(Exercise) What is the best choice for p_t ? Note that the best p_k for each particular (i, j) might be different. But it takes $O(n^3)$ time if we make this best choice for each i, j. Instead, we will make the choice of p_k the best for the overall variance; we will see that this can be worked out cleanly.

Notation: Let $||\mathbf{A}||_F^2 = \sum_{i,j} \mathbf{A}_{ij}^2$.

Lemma 4. *$p_k = \frac{|\mathbf{A}^{(k)}||\mathbf{B}_{(k)}|}{\sum_{l=1}^n |\mathbf{A}^{(l)}||\mathbf{B}_{(l)}|}$ for $k = 1, \ldots, n$, minimizes $E\left(||\mathbf{AB} - \mathbf{CR}||_F^2\right)$. More specifically, $E\left(||\mathbf{AB} - \mathbf{CR}||_F^2\right) \leq \frac{1}{c}(\sum_{k=1} n|\mathbf{A^k}||\mathbf{B_k}|)^2 - \frac{1}{c}||\mathbf{AB}||_F^2$.*

Proof. $E\left(\sum_{i,j}(\mathbf{AB} - \mathbf{CR})_{ij}^2\right) = \sum_{ij} E(\mathbf{AB} - \mathbf{CR})_{ij}^2)$ (using linearity of expectations.)

$$
\begin{aligned}
&= \sum_{ij} Var[\mathbf{CR}_{ij}] \\
&= \frac{1}{c}\sum_{ij}\sum_{k}\frac{\mathbf{A}_{ik}^2\mathbf{B}_{kj}^2}{p_k} - \frac{1}{c}||\mathbf{AB}||_F^2 \\
&= \frac{1}{c}\sum_{k}\frac{1}{p_k}\sum_{ij}\mathbf{A}_{ik}^2\mathbf{B}_{kj}^2 - \frac{1}{c}||\mathbf{AB}||_F^2 \\
&= \frac{1}{c}\sum_{k}\frac{1}{p_k}\left(\sum_{i}\mathbf{A}_{ik}^2\right)\left(\sum_{j}\mathbf{B}_{kj}^2\right) - \frac{1}{c}||\mathbf{AB}||_F^2 \\
&= \frac{1}{c}\sum_{k}\frac{1}{p_k}|\mathbf{A}^{(k)}|^2|\mathbf{B}_{(k)}|^2 - \frac{1}{c}||\mathbf{AB}||_F^2
\end{aligned}
$$

Now, $p_k = \frac{|\mathbf{A}^{(k)}||\mathbf{B}_{(k)}|}{\sum_{l=1}^n |\mathbf{A}^{(l)}||\mathbf{B}_{(l)}|}$ for $k = 1, \ldots, n$ is the best choice of probability (by exercise). With this choice we get the lemma. $\square$

To get a feeling of the case in which we have large error, let $\mathbf{B} = \mathbf{A^T}$. If $\mathbf{A}$ is an object with n features. $\mathbf{AA^T}$ measures the similarity between them. In this case the p_k's turn out to be $p_k = \frac{|\mathbf{A^K}|^2}{||\mathbf{A}||_F^2}$. $E(||\mathbf{AA^T} - \mathbf{CC^T}||_F^2) = \frac{1}{c}||\mathbf{A}||_F^2 - \frac{1}{c}||\mathbf{AA^T}||_F^2$. If $\mathbf{AA^T}$ has a lot of cancellation, then $||\mathbf{AA^T}||_F$ maybe $\leq \varepsilon||\mathbf{A}||_F^2$, leading to large error in the result.