Lecture 3 — January 14th

Lecturer: Mert Pilanci *Scribe: Eric Luxenberg*

3.1 Probing the Error

Given our approximate matrix multiplication algorithm, we have $AB \approx CR$, so we are interested in the error,

$$\Delta = AB - CR$$

A natural way to quantity the error is via its Frobenius norm, $\|\Delta\|_F$. We note that

$$\|\Delta\|_F^2 = \text{tr}(\Delta^T \Delta)$$

which motivates estimating the trace. The trace is a linear computation if we have the matrix in memory; the sense in which we are interested in estimating the trace is when we do not have the matrix available.

Example 1. Imagine wanting to know $\text{tr}(B^3)$, but being limited in compute so as to not actual compute $B \cdot B \cdot B$.

Matrix multiplication will be $O(n^3)$. There exist methods for computing (approximately) $\text{tr}(f(M))$ faster than cubic time. Namely, in our case, we want to analyze $\|\Delta\|_F$ without actually computing AB.

3.1.1 Trace Estimation

Let B be an $n \times n$ symmetric matrix. Let $u_1, ..., u_n$ be n I.I.D. 0 mean, σ^2 variance random variables. We will estimate $\text{tr}(B) \approx u^T B u$. First, we check that our estimate is unbiased.

Lemma 1. Trace estimate is unbiased with $\sigma^2 = 1$

$$\mathbb{E}[u^T B u] = \mathbb{E}[\sum_{i,j} u_i B_{ij} u_j] = \sum_{i,j} B_{ij} \delta_{ij} \sigma^2 = \sigma^2 \sum_i B_{ii} = \sigma^2 \,\text{tr}(B)$$

where we used that $\mathbb{E}[u_i u_j] = 0$ when $i \neq j$, since the components are independent, and $\mathbb{E}[u_i^2] = \sigma^2$ since the mean is zero. We see the estimate is unbiased then the variance is one.

Now, we analyze the variance of this trace estimator.

Lemma 2. Variance of trace estimator

$$\text{Var}(u^T B u) = \mathbb{E}[(u^T B u)^2] - \mathbb{E}[u^T u]^2$$

We know the latter term is $\text{tr}(B)$ when $\sigma^2 = 1$. We consider the first term.

$$\mathbb{E}[(u^T Bu)^2] = \mathbb{E}\left[\sum_{ij} u_i B_{ij} u_j \sum_{rs} u_r B_{rs} u_s\right]$$

We will now break down this expression into cases, based on the observation that $E[u_a u_b] = 0$ whenever $a \neq b$. Namely, we consider the four ways to have pairs of indexes being equal, expanding the above expression as

$$\sum_{i=j=r=s} \mathbb{E}[u_i^4]B_{ii}^2 + \sum_{i=j,r=s,r\neq i} \mathbb{E}[u_i^2]\mathbb{E}[u_r^2]B_{ii}B_{rr} + \sum_{i\neq j,i=r,s=j} \mathbb{E}[u_i^2]\mathbb{E}[u_j^2]B_{ij}^2 + \sum_{i\neq j,i=s,r=j} \mathbb{E}[u_r^2]\mathbb{E}[u_s^2]B_{rs}^2$$

$$= 2\sigma^4 \sum_{i\neq j} B_{ij}^2 + (\mathbb{E}[u^4] - \sigma^4)\sum_i B_{ii}^2$$

Now, we wish to choose the distribution for u satisfying mean zero and variance 1 (so that the estimate is unbiased) which minimizes the variance. We note that from our above expression for the variance, minimizing the variance corresponds to minimizing $\mathbb{E}[u^4] - \sigma^4$.

Note that $\text{Var}(u^2) = \mathbb{E}[u^4] - \mathbb{E}[u^2]^2 = \mathbb{E}[u^4] - \sigma^4$, so we are actually trying to choose the minimum variance mean 0 variance 1 random variable. But note that choosing U to be ± 1 with probability $1/2$ on each outcome leads to a deterministic random variable U^2, which has variance 0 and hence minimizes our desired expression. Choosing u as such (called a Rademacher random variable) is known as the "Hutchinson Trace Estimator."

3.1.2 Probing the errror, cont.

Using this trace estimator, we see that we can approximately compute

$$\|AB - CR\|_F^2 \approx \|(AB - CR)u\|_F^2$$

where we only need matrix-vector products from AB, not all of AB. Namely, we can first compute Bu, and then $A(Bu)$. This computation will be $O(n^2)$, vs $O(n^3)$ for AB.

The catch is that we do need to go back to the data to collect the matrix-vector products. One important additional observation is that the Rademacher distribution is particularly convenient since we dont need to multiply; we can just swap sign bits and add.

3.2 Sampling/Sketching Formalism

We define a sampling matrix $\hat{S}$ with $\hat{S}_{ij} = \begin{cases} 1 & \text{if the i'th column is chosen in the j'th trial} \\ 0 \end{cases}$

Define a diagonal re-weighting matrix D with $D_{tt} = \frac{1}{\sqrt{mp_{i_t}}}$ Then in our approximate matrix multiplication, we have

$$AB \approx CR$$

where

$$C = A\hat{S}D, R = D\hat{S}^T B$$

Letting $S = D\hat{S}^T$, we have

$$CR = A\hat{S}DD\hat{S}^T B = AS^T SB$$

This connects our approximate matrix multiplication algorithm to our analysis in the sequel under the lens of random dimension reduction and approximately orthogonal matrices.

3.3 Estimating the entry-wise error

We have considered the Frobenius norm error of our approximate matrix multiplication algorithm. But this is a someone opaque error analysis, in that it in a sense aggregates many different directions of error in one number (the Frobenius norm can be related to the summed singular values squared). A more acute kind of error analysis is via infinity-norm, i.e. the maximum entry-wise error. We let

$$\epsilon(S) = \|AS^T SB - AB\|_\infty = \max_{i,j} |(AS^T SB)_{ij} - (AB)_{ij}|$$

The .99-quantile of $\epsilon(S)$ is the tightest upper bound which holds with probability at least .99. To estimate this value, we perform the following bootstrap procedure:

```
For b= 1,...,B do
    sample m numbers with replacement from {1,...,m}
    form S_b by selecting the the respective rows of S
    compute eps_b ||AS_b'S_bB-AS'S||_inf
return 0.99-quantile of the values eps_1,...,eps_b,
(e.g. sort in increasing order and returnb 0.99B -th value)
```

Statistical theory shows that $\epsilon(S) \approx \frac{\kappa}{\sqrt{m}}$, and so given a bootstrap estimate of the error at some m_0, we can extrapolate the error for $m > m_0$ as

$$\frac{\sqrt{m_0}}{\sqrt{m}}\hat{\epsilon}(S)$$